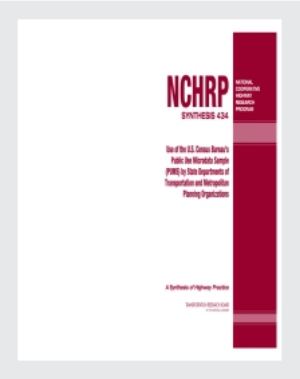


Use of the U.S. Census Bureau's Public Use Microdata Sample (PUMS) by State Departments of Transportation and Metropolitan Planning Organizations

DETAILS

73 pages | 8.5 x 11 | PAPERBACK

ISBN 978-0-309-22365-2 | DOI 10.17226/22772

AUTHORS

Tierney, Kevin F.

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

NCHRP SYNTHESIS 434

Use of the U.S. Census Bureau's Public Use Microdata Sample (PUMS) by State Departments of Transportation and Metropolitan Planning Organizations

A Synthesis of Highway Practice

CONSULTANT
KEVIN F. TIERNEY
Needham, Massachusetts

SUBSCRIBER CATEGORIES
Data and Information Technology • Highways • Pedestrians and Bicyclists • Planning and Forecasting • Public Transportation • Society

Research Sponsored by the American Association of State Highway and Transportation Officials
in Cooperation with the Federal Highway Administration

TRANSPORTATION RESEARCH BOARD

WASHINGTON, D.C.
2012
www.TRB.org

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Systematic, well-designed research provides the most effective approach to the solution of many problems facing highway administrators and engineers. Often, highway problems are of local interest and can best be studied by highway departments individually or in cooperation with their state universities and others. However, the accelerating growth of highway transportation develops increasingly complex problems of wide interest to highway authorities. These problems are best studied through a coordinated program of cooperative research.

In recognition of these needs, the highway administrators of the American Association of State Highway and Transportation Officials initiated in 1962 an objective national highway research program employing modern scientific techniques. This program is supported on a continuing basis by funds from participating member states of the Association and it receives the full cooperation and support of the Federal Highway Administration, United States Department of Transportation.

The Transportation Research Board of the National Research Council was requested by the Association to administer the research program because of the Board's recognized objectivity and understanding of modern research practices. The Board is uniquely suited for this purpose as it maintains an extensive committee structure from which authorities on any highway transportation subject may be drawn; it possesses avenues of communication and cooperation with federal, state, and local governmental agencies, universities, and industry; its relationship to the National Research Council is an insurance of objectivity; it maintains a full-time research correlation staff of specialists in highway transportation matters to bring the findings of research directly to those who are in a position to use them.

The program is developed on the basis of research needs identified by chief administrators of the highway and transportation departments and by committees of AASHTO. Each year, specific areas of research needs to be included in the program are proposed to the National Research Council and the Board by the American Association of State Highway and Transportation Officials. Research projects to fulfill these needs are defined by the Board, and qualified research agencies are selected from those that have submitted proposals. Administration and surveillance of research contracts are the responsibilities of the National Research Council and the Transportation Research Board.

The needs for highway research are many, and the National Cooperative Highway Research Program can make significant contributions to the solution of highway transportation problems of mutual concern to many responsible groups. The program, however, is intended to complement rather than to substitute for or duplicate other highway research programs.

NOTE: The Transportation Research Board of the National Academies, the National Research Council, the Federal Highway Administration, the American Association of State Highway and Transportation Officials, and the individual states participating in the National Cooperative Highway Research Program do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of this report.

NCHRP SYNTHESIS 434

Project 20-05 (Topic 42-02)
ISSN 0547-5570
ISBN 978-0-309-22365-2
Library of Congress Control No. 2012934392

COPYRIGHT INFORMATION

NOTICE

The project that is the subject of this report was a part of the National Cooperative Highway Research Program conducted by the Transportation Research Board with the approval of the Governing Board of the National Research Council. Such approval reflects the Governing Board's judgment that the program concerned is of national importance and appropriate with respect to both the purposes and resources of the National Research Council.

The members of the technical committee selected to monitor this project and to review this report were chosen for recognized scholarly competence and with due consideration for the balance of disciplines appropriate to the project. The opinions and conclusions expressed or implied are those of the research agency that performed the research, and, while they have been accepted as appropriate by the technical committee, they are not necessarily those of the Transportation Research Board, the National Research Council, the American Association of State Highway and Transportation Officials, or the Federal Highway Administration of the U.S. Department of Transportation.

Each report is reviewed and accepted for publication by the technical committee according to procedures established and monitored by the Transportation Research Board Executive Committee and the Governing Board of the National Research Council.

Published reports of the

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

are available from:

Transportation Research Board
Business Office
500 Fifth Street, NW
Washington, DC 20001

and can be ordered through the Internet at:
http://www.national-academies.org/trb/bookstore

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. On the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, on its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

The **Transportation Research Board** is one of six major divisions of the National Research Council. The mission of the Transportation Research Board is to provide leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied activities annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation. **www.TRB.org**

www.national-academies.org

FOREWORD

Highway administrators, engineers, and researchers often face problems for which information already exists, either in documented form or as undocumented experience and practice. This information may be fragmented, scattered, and unevaluated. As a consequence, full knowledge of what has been learned about a problem may not be brought to bear on its solution. Costly research findings may go unused, valuable experience may be overlooked, and due consideration may not be given to recommended practices for solving or alleviating the problem.

There is information on nearly every subject of concern to highway administrators and engineers. Much of it derives from research or from the work of practitioners faced with problems in their day-to-day work. To provide a systematic means for assembling and evaluating such useful information and to make it available to the entire highway community, the American Association of State Highway and Transportation Officials—through the mechanism of the National Cooperative Highway Research Program—authorized the Transportation Research Board to undertake a continuing study. This study, NCHRP Project 20-5, "Synthesis of Information Related to Highway Problems," searches out and synthesizes useful knowledge from all available sources and prepares concise, documented reports on specific topics. Reports from this endeavor constitute an NCHRP report series, *Synthesis of Highway Practice*.

This synthesis series reports on current knowledge and practice, in a compact format, without the detailed directions usually found in handbooks or design manuals. Each report in the series provides a compendium of the best knowledge available on those measures found to be the most successful in resolving specific problems.

PREFACE

By Jon M. Williams
Program Director
Transportation
Research Board

Census microdata are the confidential records of specific individuals and housing units from whom Decennial Census or American Community Survey responses have been obtained. The U.S. Census Bureau also draws a sample from the full set of microdata and makes these sampled records available in the Public Use Microdata Sample (PUMS) data products, so that users can develop their own tabulations. These data are being used by state departments of transportation (DOTs) and metropolitan planning organizations (MPOs) for studies, such as analyses of the commuting characteristics of population subgroups, and for supporting travel demand model and land use models.

Information for this study of PUMS use was gathered by literature review, survey of selected state DOTs and MPOs, and in-depth interviews.

Kevin F. Tierney, Needham, Massachusetts, collected and synthesized the information and wrote the report. The members of the topic panel are acknowledged on the preceding page. This synthesis is an immediately useful document that records the practices that were acceptable within the limitations of the knowledge available at the time of its preparation. As progress in research and practice continues, new knowledge will be added to that now at hand.

CONTENTS

Note: Many of the photographs, figures, and tables in this report have been converted from color to grayscale for printing. The electronic version of the report (posted on the web at www.trb.org) retains the color versions.

USE OF THE U.S. CENSUS BUREAU'S PUBLIC USE MICRODATA SAMPLE (PUMS) BY STATE DEPARTMENTS OF TRANSPORTATION AND METROPOLITAN PLANNING ORGANIZATIONS

SUMMARY

Transportation planners in some regions are using Census Public Use Microdata Sample (PUMS) data as essential inputs to mission-critical analyses, but planners at most agencies are not familiar with these data or their benefits. This synthesis seeks to describe how transportation planners are using the PUMS data and to serve as a reference for transportation planners who may be able to exploit these data.

PUMS data are somewhat unusual for Census data in that they are not tabulations of data summarized at a specified geographic area. Instead, they are a sample of the actual data records of the information collected in the American Community Survey and previously in the Decennial Census. The PUMS records are subjected to data disclosure avoidance techniques to protect respondents' confidentiality. Most notably, the most precise geographic areas allowed for PUMS data must have at least 100,000 residents.

Information for this study was selected by a literature review, web-based survey, and in-depth interviews. Respondents for the survey included 37 state departments of transportation (DOTs), a 71% response rate; 23 large metropolitan planning agencies (MPOs), a 55% response rate; and 25 small MPOs, a 33% response rate.

Slightly more than one-third of the state DOTs contacted in this synthesis effort are regular or occasional users of PUMS data. About two-thirds of the large MPOs that participated in the synthesis use PUMS data, most on a regular basis. However, few small and medium MPOs that participated in this review use these data. This usage is substantially below the usage of many other Census data products and other similar databases. To some extent, these differences in usage reflect the fact that the PUMS data set is a rather specialized, niche data product. However, the most common reason that nonusers gave for not using the PUMS data was their lack of familiarity with these data. Roughly half of nonusers contacted are not completely aware of what the PUMS data are. A perceived lack of technical understanding of PUMS also plays an important role in agencies choosing not to use the data.

In contrast, PUMS data users generally rate their importance highly and their levels of satisfaction with the data relatively highly, as well. Because the PUMS data include full records with a wide range of household and person data items, data users are able to cross-tabulate and explore relationships between different data combinations than the Census Bureau can provide in its standard products or than FHWA and AASHTO can provide in the Census Transportation Planning Products (CTPP) data tables. Transportation planners and researchers have found PUMS to be especially useful for the following:

- **Cross-tabulations of variables not readily available from CTPP** – The Census and CTPP tables often enable transportation planners to easily locate information needed to support planning applications, but on occasion, analysis requires combining population characteristics that are not included in those tabulations. Often, these analyses

are looking at special subpopulations (e.g., members of ethnic groups, people of certain ancestries, group quarters residents, bicycle commuters) that can be separated using the PUMS data.
- **Cross-tabulations of variables in CTPP but with more currency** – Because PUMS data are available on an ongoing basis and the CTPP are available periodically, planners can use the PUMS data to create more up-to-date CTPP-like data tables, albeit with less precision in the estimates and less geographic detail.
- **Disaggregate analyses** – Planners and modelers frequently require household-level or person-level (disaggregate) data to develop models of the interrelationships between household and person characteristics. The microdata represented by the PUMS data allow users to evaluate variable relationships at the housing unit and person level.
- **Comparisons of different regions** – Because the PUMS data series provides common data sets for all regions of the country, and the Census Bureau provides the same attention to detail in its data collection efforts, the PUMS data are particularly useful for interregional comparisons and national analyses.
- **Comparisons over time** – PUMS data sets exist for each of the Decennial Census data collection efforts and for each year of American Community Survey (ACS) implementation, so the data are commonly used to track changes in housing and person characteristics and changes in the interrelationships between these characteristics over time. Minor changes in the variables and reporting levels make these comparisons more difficult, but planners and researchers are only beginning to explore the utility of annual PUMS data.
- **Validation of other data sources** – PUMS data can be used to check calculations and predictions made using other data sources, such as travel surveys, demographic estimates, and modeling results.

These tabulations and analyses frequently support specific issues and studies, such as analyses of the commuting characteristics of specific population subgroups or the demographic characteristics of commuters by mode.

The PUMS data are also used to support travel surveys and travel demand models. PUMS data are used in the planning, design, expansion, and validation of household travel surveys, and also provide input data for the development and validation of state-of-practice travel demand model subcomponents. In recent years, PUMS data have been used to support the development of household composition and auto availability submodels, trip generation models, and external trip models. PUMS data also provide base year information for travel model validation and checking.

Finally, as the developers of advanced travel demand models and integrated transportation land use models are relying on microsimulation techniques to a greater extent, the usage and importance of PUMS data have increased. All of the activity-based travel demand models that have been or are being developed in the United States rely on PUMS data as a key input into the population synthesis module of their model systems. In addition, several of the most widely used land-use modeling systems use PUMS data for the same reason.

Because of the unique nature of the PUMS data, there are no potential alternative data products. Transportation planners who use PUMS data are likely to continue to do so until underlying data needs change substantially, and over time, as more agencies increase their demand modeling capabilities, more transportation planners will need to become PUMS experts.

To address the needs of both expert users and those who lack familiarity with the PUMS data, this synthesis suggests several further research activities.

More research is needed to determine the best ways for transportation planning agencies to seek out and take advantage of the Census Bureau PUMS training materials and documenta-

tion. The transportation planning community may benefit from discussions of PUMS data quality issues and benefits in our two-way dialogue with the Census Bureau staff. Through the CTPP process and the advocacy of FHWA data specialists, the Census Bureau has developed a richer understanding of transportation planning data uses, and the planning community has a better understanding of the Census Bureau's expertise and constraints.

In addition, the data user group will benefit from further research on how best to promote and disseminate technical research in areas related to the PUMS. The proposed research includes the following:

- Research on the PUMS data series development to better understand whether the standard PUMS data products could be improved without affecting Census Bureau disclosure requirements;
- Development of transportation PUMS user resources that will promote the sharing of PUMS-based research, analysis tools, and data processing computer code, especially among MPOs with less technical staff;
- Continued research on technical and methodological issues related to population synthesis;
- New research on how the Census Bureau migration to the ACS affects the implementation and design of population synthesis models; and
- New research on how the introduction of new synthetic CTPP tables will affect the implementation and design of population synthesis models.

CHAPTER ONE

INTRODUCTION

PURPOSE OF THE SYNTHESIS

Transportation planners make extensive use of standard Census Bureau data tabulations and summaries from the Decennial Census and American Community Survey (ACS). Transportation users' high level of interest in and usage of these data has led AASHTO to sponsor the development and dissemination of the Census Transportation Planning Products (CTPP) special tabulations. Transportation planners may also benefit further from the use of other, less commonly used, Census data products.

In recent years, the Census Public Use Microdata Sample (PUMS) data developed as part of the Decennial Census and ACS data programs appear to be becoming a more common data set for transportation planners. Planners have needed to find new data sources to analyze population subgroups, and the PUMS data can help in this regard. In addition, as transportation planning agencies are implementing more sophisticated travel demand models—many of which rely on the microsimulation of synthesized household and person travel behavior—there is increased need for disaggregate data such as those provided by the PUMS data sets.

The Census Bureau provides excellent documentation on the use of PUMS, and both the Census Bureau and some universities offer users different ways to access the PUMS data. However, because the primary users of PUMS data are in the academic research community, professional transportation planners often are not as aware of these data as they are of other Census data resources. This makes it more difficult for members of the transportation community to learn from each other or to advise those responsible for designing usable data products for the transportation community. Compounding this information void, the recent change to the ACS from the Census Long Form Survey has introduced new methodological issues and concerns that are not well understood.

The purpose of this synthesis is to better define—

- Who the transportation users of the PUMS data are,
- To what extent and for what purposes these analysts are using the PUMS data, and
- The benefits, limitations, and data issues these analysts have encountered in their PUMS usage.

This synthesis effort seeks to describe an array of transportation-related PUMS data uses so that transportation planners can better understand whether they can effectively take advantage of the PUMS data in future analyses. Finally, the synthesis seeks to identify potential ways to improve transportation planners' ability to use PUMS data.

STUDY METHODOLOGY

The synthesis report included three tracks:

- Review of published and unpublished documentation on PUMS usage by transportation planners,
- In-depth interviews with transportation planners that use PUMS, and
- Web-based scan of Census data users within transportation agencies.

Literature Review

The data assembly process began with a review of published literature for which PUMS data played a role. Transportation planning papers and reports that describe the use of PUMS data were gathered and summarized throughout the synthesis effort. These research and planning efforts were classified into analysis categories to help organize the synthesis.

In-Depth Interview Contacts

Many PUMS data users were interviewed by telephone. The respondents included those identified from published research, those identified in the state department of transportation (DOT) and metropolitan planning organization (MPO) survey discussed later, and those identified in other telephone interviews. These interviews comprised detailed discussions about how the PUMS data had been used and about researchers' views of the PUMS data. Information and respondent views on the likely transferability and applicability of the analyses by other analysts were also sought.

In-depth interviews were conducted with members of the following groups:

- State DOT and MPO survey respondents who had indicated PUMS usage,

- Researchers who have published reports and papers describing in part the use of PUMS data for transportation planning,
- Academic researchers in the transportation planning field,
- TRB data committee members, and
- Transportation planning consultants and market research consultants who have worked with agencies on efforts involving the use of PUMS data.

In-Depth Interview Content

The in-depth interviews primarily involved open-ended discussions. The informal interview discussion guide included the following:

- Introductions and explanation of the PUMS synthesis effort, and how the respondents were identified as potential participants;
- Broad summary of the current understanding of the work that was conducted and how PUMS data were used (if specific PUMS-based efforts had been identified for the respondent);
- A summary of the work and any available documentation (if no advance knowledge of the specific PUMS-based efforts had been obtained);
- Assessment of whether the effort used Decennial PUMS, single-year ACS PUMS, multiyear ACS PUMS, or a combination; as well as participants' views on how the ACS migration might affect the analyses performed;
- Source of PUMS data (directly from Census Bureau, Integrated Public Use Microdata Series, other);
- Software/programming languages used for data processing/analyses;
- How other data sources were used in conjunction with PUMS data;
- Questions about any positive aspects, problems, or issues with using PUMS data, including the following:
  - Ease of data access
  - Ease of data manipulation and processing
  - Limitations of PUMS geography
  - Limitations of sampling procedures
  - Limitations of data disclosure procedures
  - Adequacy of documentation
  - "Wish list" of improvements to PUMS data
- Possible future plans for analyses;
- Transferability and adaptability of analyses;
- Opportunities for collaboration, knowledge sharing, and improvement of PUMS data set and dissemination;
- Any other uses of PUMS data by participants, and repeat discussion topics for each use of the data; and
- Questions about others' uses of PUMS data, and contact information for these other potential participants.

Web-Based Survey Scan

The web survey scan sought to shed some light on the breadth of PUMS usage by planners at state DOTs and MPOs. The survey was designed using a web interviewing package provided by TRB. Potential survey respondents were contacted by email and invited to go to a website to complete the survey. Survey respondents who were using the PUMS data were asked to provide additional information through more in-depth interviews on their specific PUMS uses.

Survey Scan Sample

The survey was intended to provide data from a range of planning agencies so that informal inferences could be made about the extent and nature of PUMS usage. Because the survey may not have reached all of the right people within the responding agencies and because survey nonresponse follow-up contacts were not performed for the entire survey sample, the survey results should not be considered statistically robust. The survey results and findings probably reflect the true breadth of PUMS usage by the agencies, but without the statistical precision that more complete surveys sometimes attain. The survey universe included the 52 state DOTs (includes District of Columbia and Puerto Rico) on the AASHTO Standing Committee on Planning Board and the 381 MPOs. Because the usage of PUMS data was thought likely to vary by the size and sophistication of the agencies, the MPOs were stratified by metropolitan area size.

All the state DOTs and larger MPOs were invited to participate. These agencies are the most likely to be active PUMS data users, and therefore were more likely to have a wider range of survey responses. The sample of smaller agencies was randomly selected from the FHWA's list of current MPOs.

Survey Scan Method

The key data collection steps included the following:

- Pretest with panel members by sending them the survey invitation e-mails with the survey link. Panel members had the opportunity to complete the survey just as the actual respondents would, and to provide comments. The survey instrument and procedures were edited based on these comments.
- Try to identify best contacts at sampled agencies through consultation with Panel members and agency website review.
- Send NCHRP staff e-mail survey invitations to state DOT and MPO contacts. The invitations allowed potential respondents to click on a hyperlink to go the survey. The web survey software allows for the internal tracking of individual respondents, so it was possible to determine whether an invitee had responded.

This feature let respondents partially complete the survey, save it, and return at a later time. This allowed respondents to consult with colleagues and to locate project documents.
- The survey scan remained open for about 6 weeks. Several days after the initial invitations had been sent, NCHRP sent a reminder email with the link to nonrespondents. After several more days, a sample of nonrespondents were called and asked the survey questions over the phone.
- At the end of the survey scan period, respondents who indicated that they would be willing to provide more details about their specific PUMS analyses were contacted.

Survey Scan Content

Appendix A shows a draft script for the web survey used in the scan. The web survey software estimate for average completion time was 13 minutes.

ORGANIZATION OF THE REPORT

The remainder of this report describes the synthesis effort. Chapter two provides an overview of the PUMS data for readers who are not experts. As discussed later, transportation planners have a relatively low level of familiarity with the PUMS data source. This section describes the key issues related to the unique features of the PUMS data.

Chapter three describes transportation planners' familiarity with and usage levels of the PUMS data. This chapter summarizes the results of the web-based survey scan of transportation agencies. The scan sought information on who is using PUMS data and why, as well as who is not using PUMS and why not. It showed that agencies that use the PUMS data generally find them to be important and useful, but that a large proportion of agencies are generally unfamiliar with PUMS and unaware of whether the data might be of value to them.

Chapter four summarizes several ways that transportation planners are taking advantage of the PUMS data to support their analyses. It first describes the most common general use of the PUMS data—the straightforward analyses and cross-tabulations of particular PUMS variables that support specific projects and research efforts. In many cases, the unique characteristics of the PUMS data enable these data mining efforts. Similar analyses would not be possible without the PUMS. Chapter four then discusses several ways that planners use PUMS to support travel survey efforts and travel demand modeling efforts. The PUMS data set is not a necessary source for most conventional travel demand models, but some modelers have been able to improve model components and submodels with the PUMS data. The chapter also describes the use of PUMS data in the development of advanced travel demand models and land use models. The PUMS data have become an essential input into U.S. activity-based travel demand models. These models are applied through microsimulation of household and person behavior, so the first step of the model process is the development of a synthetic population for the model area. PUMS data provide a unique type of data for developing synthetic populations and households.

Chapter five summarizes the overall findings of the review of PUMS data use by transportation planners. It makes recommendations to transportation planners and agencies on how to best take advantage of the PUMS data. It also summarizes the issues that users have identified with the PUMS data in order to identify ways that the data dissemination might be improved.

Appendix A includes the web survey scan questionnaire. Appendix B provides more technical discussion of the usage of PUMS data to support microsimulation modeling.

CHAPTER TWO

OVERVIEW OF THE PUBLIC USE MICRODATA SAMPLE DATA

DESCRIPTION OF PUBLIC USE MICRODATA SAMPLE DATA

Most Census Bureau data products are summaries and tabulations of data collected from one of the Bureau's many large-scale data collection efforts. For the Decennial Census and ACS, the Census Bureau provides single-variable tables at different prespecified geographic levels, and makes them available on the American FactFinder website (http://factfinder.census.gov). In addition, through its special tabulations programs, the Census Bureau provides cross-tabulations that are known to be of interest to specific data user groups. For instance, the CTPP, sponsored by AASHTO, are a set of Census special tabulations that support transportation planning (http://ctpp.transportation.org/Pages/default.aspx).

Because it would be impossible for the Census Bureau to anticipate every potential intervariable relationship that data users (particularly in the research community) would like to explore and to provide tabulations for them, the Census Bureau began providing PUMS data in the mid-twentieth century. The Census Bureau has provided these data for each subsequent Decennial Census and for the ACS.

Census microdata are the actual confidential records of specific individuals and housing units from which Decennial Census or ACS responses have been obtained. The Census Bureau performs analyses and summaries of the full set of microdata to develop the tabulations, reports, and special tabulations that it makes available to the public. These are the most commonly used Census Bureau products. However, the Census Bureau also draws a sample from the full set of microdata, and makes these sampled records available in the PUMS data products, so that users can develop their own tabulations.

The key advantage of the PUMS data is that data users can investigate intervariable relationships and develop cross-tabulations that are not available in tabulations provided by the Census Bureau. For instance, PUMS data users can summarize the characteristics of unemployed homeowners or compare people with specific ancestries across age categories. PUMS files allow users to separately analyze household and group quarters populations that can sometimes be intermingled in published tables. Data users can also investigate a wider range of multivariable relationships than the standard published tabulations are able to show through larger n-way tables or through population simulation (U.S. Census Bureau 2008).

The Census Bureau provides PUMS data files for the year 2000 Census and for the ACS. The year 2000 PUMS files consist of a 5 percent sample and a 1 percent sample. Beginning with the 2005 ACS, the Census Bureau has produced annual 1-year PUMS files that include 1 percent of the nation's housing unit records. Beginning with the 2006 ACS, the Census Bureau has also produced annual 1-year PUMS files that include 1 percent of the people in group quarters. The Census Bureau has combined successive 1-year PUMS files into 3-year and 5-year files, representing 3 and 5 percent of housing records and group quarters residents over the relevant period. The most recently available 5-year PUMS file (2005–2009) is based on only 4 years of group quarters data, but subsequent files will include data for all 5 years in the period (U.S. Census Bureau 2009a).

The Census Bureau has developed a PUMS handbook as part of its ACS Compass Products collection. This book, titled *What Public Use Microdata Sample (PUMS) Data Users Need to Know*, is available at http://www.census.gov/acs/www/guidance_for_data_users/handbooks/. The Census Bureau has also prepared an introductory PowerPoint presentation on PUMS, which is available at http://www.census.gov/acs/www/guidance_for_data_users/training_presentations/.

Table 1 lists PUMS data sets and where they can be found.

American Community Survey Public Use Microdata Sample versus Decennial Census Public Use Microdata Sample

The PUMS data from the ACS and the Decennial Census long form are in principle the same. They are a sample of the collected data records that have been subject to disclosure protection. However, because the data collection efforts have important conceptual differences, the resulting PUMS data files will also have important differences.

The migration to the continuous ACS from the Census long form "snapshot" survey has two significant benefits for data users. First, the ACS data that users can access are more current and therefore more relevant for analyses. Second, the

TABLE 1
AVAILABLE CENSUS BUREAU PUMS DATASETS

PUMS Data Set	Format	Database Access
ACS Five Year PUMS • 2005–2009	Downloadable files	Census ACS website http://www.census.gov/acs/www/data_documentation/pums_data/
ACS Three Year PUMS • 2007–2009 • 2006–2008 • 2005–2007	Downloadable files	Census ACS website http://www.census.gov/acs/www/data_documentation/pums_data/
ACS One Year PUMS • 2009, 2008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000	Downloadable files	Census ACS website http://www.census.gov/acs/www/data_documentation/pums_data/
ACS Test Site PUMS • Florida, New York, Oregon (1996–1998) • Nebraska, Ohio, Texas (1997–1998) • South Carolina (1998)	Downloadable files	Census ACS website http://www.census.gov/acs/www/data_documentation/pums_data/
2000 Decennial Census Five-Percent PUMS	Downloadable files	Census 2000 website http://www.census.gov/census2000/PUMS5.html
2000 Decennial Census One-Percent PUMS	Downloadable files	Census 2000 website http://www.census.gov/census2000/PUMS.html
1990 Decennial Census Five-Percent PUMS and One-Percent PUMS	DVD and CD-ROM	Purchase from Census Bureau http://www.census.gov/mp/www/cat/decennial_census_1990/1990_census_of_population_and_housing_public_use_microdata_samples_pums_5_and_1.html
1980 Decennial Census Five-Percent PUMS	CD-ROM	Purchase from Census Bureau http://www.census.gov/mp/www/cat/decennial_census_1980/1980_census_of_population_and_housing_public_use_microdata_pums_5.html
Pre-1980 PUMS	Microfilm	Accessible at National Archives sites http://www.archives.gov/research/census/

Source: Compiled from U.S. Census Bureau website pages (April 2011).

ACS data, on a record-by-record basis, are likely to be more accurate because the continuous data collection program enables better quality control and maintenance procedures to be enacted. These improvements are two of the important reasons one would expect different results between the data collection efforts (Cambridge Systematics et al. 2007).

The primary drawback to the switch to the ACS is the reduction in the sample size and lower (unweighted) response rates. The Decennial Census long form data collection included more than 15 percent of the households and group quarters residents, but each ACS year includes only about 2.5% of these groups. Even with multiyear accumulation of ACS data, the precision of estimates is reduced, and the need for measuring and reporting confidence intervals is increased. The ACS PUMS sample sizes have been set so that a 5% PUMS sample will be available for a 5-year accumulation of data. Rather than having a 5 percent sample of households and group quarters residents for a single year, as with the Decennial Census data, ACS PUMS data users will need to work with a smaller 1% sample of data for the year or with data collected over a 3-year or 5-year period to get a larger sample of records (Cambridge Systematics et al. 2007).

Significant conceptual differences also affect data comparisons. It is important to remember that ACS estimates are period estimates collected continuously throughout the year. The 2000 Decennial Census data are collected for a specific date. This difference necessitates some conceptual differences between ACS and decennial data (Cambridge Systematics et al. 2007; U.S. Census Bureau 2009b):

- Residence rules: The Decennial Census assigns respondents to residences based on their "usual residence" as of April 1, 2000; ACS assigns respondents to residences based on their "current residence." Appreciable differences between the methods will occur for areas with highly seasonal populations. The Decennial Census data capture one season of 1 year, without information on the rest of the year. ACS captures an average residency/vacancy condition over the course of a year, which may not exist at any one point in time.
- Reference periods: Retrospective questions that ask about a previous period of time ("In the past 12 months…," "Last week…," etc.) are affected by when the question is posed. For the Decennial Census, the responses were collected in March through June. For the ACS, the responses are collected equally throughout the year. Thus, if char-

acteristics (e.g., means of transportation to work, school enrollment, time leaving for work) change throughout the year, the surveys will measure different things.

Nature of Microdata

Figure 1 shows a Census Bureau illustration of the difference between summary data that are commonly accessed by Decennial Census and ACS data users versus PUMS data (U.S. Census Bureau 2009a). The top table of the figure shows how the summary data might look. Summary information is obtained for specific geographic areas by combining information from specific data records (microdata). The second and third tables show the layout of the PUMS data records (one table for household records and one for person-based records).

The PUMS file records represent only a sample of all the records for the entire population, so weights are needed to expand PUMS analyses to the overall population. The Decennial Census and ACS data collection efforts rely on a complex sampling strategy, so the weights will vary from record to record. The Census Bureau provides basic tabulations of weighted characteristics from the ACS PUMS that researchers can use to verify the accuracy of their data files (U.S. Census Bureau 2009a).

Public Use Microdata Sample Data File Subject Areas

The PUMS data consist of two linked record types: housing unit records and person records. The group quarters data consist of the person records and pseudo-housing unit records with zero housing unit weights. A variable in the data set allows users to link the housing unit and person records so that they can explore interrelationships between different household and person characteristics and develop "do-it-yourself" cross-tabulations with different combinations of the subjects.

The PUMS data files contain a nearly full set of data collected from the 2000 Decennial Census and the ongoing ACS effort. Tables 2 and 3 show the list of PUMS subjects for the household and person PUMS files. The slight differences in the PUMS files over the years reflect the changes that have been made to the ACS questions and response categories over time. PUMS data users who use more than one file also need to be aware of differences in the questions and answer categories even when the same subjects are covered in the files. Full data dictionaries for the Census PUMS data files are available from the Census website: http://www.census.gov/acs/www/data_documentation/pums_documentation/.

Example Summary Product

Geography	Males	Females	Median age	Occupied housing units	Owner occupied units	Renter occupied units
State 1	7,345,968	7,952,709	35.9	5,689,354	3,005,973	2,683,381
County A	45,678	49,852	33.5	40,678	15,961	24,717

Example Public Use Microdata Sample Population Records

Household ID	Person ID	PUMA	Relationship	Sex	Educational attainment	Employment status
105	1	00100	Householder	Female	Bachelors	Working
105	2	00100	Spouse	Male	Masters	Working
105	3	00100	Child	Male	Some High school	N/A

Example Public Use Microdata Sample Housing Unit Household Records

Household ID	PUMA	Tenure	Rooms	Bedrooms	Value	Contract rent
105	00100	Owned	8	3	236,500	N/a
106	00100	Rented	3	1	N/A	1,250

NOTES: In the actual PUMS files, all variables, such as tenure and relationship, are represented by numeric codes rather than descriptive text. {Artificial data}

FIGURE 1 Census Bureau conceptual comparison of summary products and public use microdata sample products. *Source*: U.S. Census Bureau (2009a).

TABLE 2
HOUSEHOLD-BASED SUBJECTS INCLUDED IN CENSUS PUBLIC USE MICRODATA SAMPLE FILES

2000 PUMS (1 percent and 5 percent samples)	2004–2008 ACS PUMS and Multiyear PUMS Containing Those Years	2009–? ACS PUMS
Agricultural sales	Agricultural sales	Agricultural sales
		Bathtub or shower
Bedrooms	Bedrooms	Bedrooms
Building size		
Commercial use	Commercial use	Commercial use
Condominium fee	Condominium fee	Condominium fee
Cost of utilities and fuels	Cost of utilities and fuels	Cost of utilities and fuels
Family income	Family income	Family income
Farm/nonfarm		
Fire, hazard, and flood insurance	Fire, hazard, and flood insurance	Fire, hazard, and flood insurance
	Food stamps	Food stamps
Fuels used	Fuels used	Fuels used
Grandparent/grandchild	Grandparent/grandchild	Grandparent/grandchild
Household and family type	Household and family type	Household and family type
Household income	Household income	Household income
Household language	Household language	
	Housing costs	Housing costs
Kitchen facilities	Kitchen facilities	
Linguistic isolation	Linguistic isolation	Linguistic isolation
Lot size	Lot size	Lot size
Meals included in rent	Meals included in rent	Meals included in rent
Mobile home costs	Mobile home costs	Mobile home costs
Mortgage payment	Mortgage payment	Mortgage payments
		Multigenerational household
Plumbing facilities	Plumbing facilities	
Presence and age of own children	Presence and age of own children	Presence and age of own children
Property taxes	Property taxes	Property taxes
Property value	Property value	Property value
		Refrigerator
Rent	Rent	Rent
Residence state	Residence state	Residence state
Rooms	Rooms	Rooms
		Running water
		Sink
		Stove
Subfamilies	Subfamilies	Subfamilies
Telephone in unit	Telephone in unit	Telephone in unit
Tenure in home	Tenure in home	Tenure in home
		Toilet
Units in structure	Units in structure	Units in structure
	Unmarried partner	Unmarried partner
Vacancy status	Vacancy status	Vacancy status
Vehicles available	Vehicles available	Vehicles available
Work	Work	Work
Year householder moved into unit	Year householder moved into unit	Year householder moved into unit
Year structure built	Year structure built	Year structure built

Sources: http://www.census.gov/acs/www/data_documentation/pums_documentation/ and U.S. Census Bureau (2003).

TABLE 3
PERSON-BASED SUBJECTS INCLUDED IN CENSUS PUBLIC USE MICRODATA SAMPLE FILES

2000 PUMS (1% and 5% samples)	2004–2008 ACS PUMS and Multiyear PUMS Containing Those Years	2009–? ACS PUMS
Ability to speak English	Ability to speak English	Ability to speak English
Age	Age	Age
Ancestry	Ancestry	Ancestry
Citizenship and naturalization	Citizenship and naturalization	Citizenship and naturalization
Class of worker	Class of worker	Class of worker
Commuting to work	Commuting to work	Commuting to work
Disability by type		
		Disability
Earnings	Earnings	
Educational attainment	Educational attainment	Educational attainment
Fertility	Fertility	Fertility
		Field of degree
Grandparent/grandchild	Grandparent/grandchild	Grandparent/grandchild
		Health insurance
Hispanic origin	Hispanic origin	Hispanic origin
Hours worked	Hours worked	Hours worked
Income	Income	
		Income by type
Industry	Industry	Industry
Language spoken at home	Language spoken at home	Language spoken at home
Last week work status	Last week work status	Last week work status
Marital status	Marital status	
		Marital status and marital history
Migration	Migration	Migration
Military	Military	Military
Mobility status	Mobility status	Mobility status
Occupation	Occupation	Occupation
Place of birth	Place of birth	Place of birth
Place of work	Place of work	Place of work
Poverty	Poverty	Poverty
Race	Race	Race
Relationship	Relationship	Relationship
School enrollment	School enrollment	
Sex	Sex	Sex
Weeks worked	Weeks worked	Weeks worked
Work	Work	Work
Year of entry	Year of entry	Year of entry

Sources: http://www.census.gov/acs/www/data_documentation/pums_documentation/ and U.S. Census Bureau (2003).

Disclosure Avoidance

Because the microdata represent complete records of actual individual Census data responses, and the Census Bureau is required by law to protect the confidentiality of respondents, the Census Bureau must take several precautions to preserve data confidentiality when publishing PUMS data (U.S. Census Bureau 2009a). These precautions include the following measures:

- PUMS data include only a sample of the data collected, rather than all the records.
- The minimum sizes of PUMS geographic reporting areas, known as Public Use Microdata Areas (PUMAs), are significantly larger than for most Census data products.
- PUMS data do not include any personally identifiable data fields, such as names or addresses.

- The Census Bureau anonymously swaps a small number of data records with other data records with similar characteristics from nearby areas.
- The Census Bureau replaces open-ended question extreme values that might allow for the identification of specific people or households with state-specific top-coded values and/or bottom-coded values.
- The Census Bureau collapses some the detail of some categorical variables.
- The Census Bureau randomly adjusts a subset of reported ages.

Public Use Microdata Sample Sampling

For the 1990 and 2000 Decennial Censuses, about 15% of households and group quarters residents were asked to complete the long form version of the questionnaire. From these responses, the Census Bureau developed two PUMS files: the 1% file (with detailed data records for 1% of the people, households, and group quarters residents) and the 5% file (with detailed data records for 5% of the people, households, and group quarters residents). The 2010 Decennial Census did not include a long form version, as this data collection was migrated to the ACS.

The ACS seeks to collect information from about 2.5% of households and group quarters residents each year, and from these responses the Census Bureau develops PUMS data files with detailed data records for 1% of the people, households, and group quarters residents. So, on average, one could expect roughly 40% of ACS data records to be included in the ACS PUMS data (after being subjected to the data disclosure limitations summarized earlier). This percentage will vary slightly owing to the sampling procedures and PUMA definitions used in the ACS. One consequence of the PUMS sampling is that neither the Census 2000 PUMS nor the ACS PUMS tabulations will exactly match the summary tables from the American FactFinder website, owing to sampling errors.

Because ACS data are being summarized in 1-, 3-, and 5-year files to allow data users to analyze smaller geographic areas by combining successive years' data, the Census Bureau is also producing 1-, 3-, and 5-year ACS PUMS data files. The multiyear PUMS data files combine the relevant 1-year PUMS files with adjustments to the weights and inflation adjustment factors. The 5-year ACS summary files include data from 12.5% of households and group quarters residents collected over the 5-year period, and the corresponding ACS PUMS data file includes records for 5% of the population.

Because PUMS data files include only a sample of data records, and because of the decrease in sample sizes, the Census Bureau advises data users to pay attention to variables' standard errors, margins of errors, and confidence intervals. The Census Bureau provides guidance on how users can calculate generalized standard errors, and also provides replicate weights for calculations of direct standard errors as part of its "PUMS Accuracy" report series: http://www.census.gov/acs/www/data_documentation/pums_documentation/.

Public Use Microdata Sample Geography

For the 2000 Decennial Census, the Census Bureau defined PUMAs for use with the 5% PUMS records and larger "super-PUMAs" for the 1% PUMS records. Each PUMA was designed to aggregate one or more counties, census tracts, minor civil divisions (MCDs), or incorporated places within a state, and each was required to contain a population of at least 100,000 people. Super-PUMAs were designed to have populations of at least 400,000 people. In addition, the Census Bureau defined Place-Of-Work PUMAs (POW-PUMAs) to provide detailed characteristics for workers and their workplaces, and Migration PUMAs (MIG-PUMAs) to provide detailed characteristics for migrants. The Census Bureau worked with state data centers (SDCs) to define PUMAs for the year 2000 Census data that had geographic components with similar characteristics.

The initial ACS PUMS files through the year 2009 have used the year 2000 PUMA definitions (with one exception, which was necessitated by the population displacement following Hurricane Katrina). The Census Bureau provides maps of Super-PUMA and PUMA geography for each state (http://www.census.gov/geo/www/maps/puma5pct.htm). Figure 2 shows one of these maps for an area of southern Ohio. GIS shapefiles of PUMAs are available from the Census Bureau TIGER/Line products website: http://www.census.gov/geo/www/tiger/index.html. In addition, the Census Bureau provides detailed equivalency files that summarize the PUMA geographic boundaries in terms of standard year 2000 Census geographic areas (http://www.census.gov/main/www/pums.html). Figure 3 shows an excerpt of the equivalency file for the southern Ohio PUMAs shown in Figure 2. Many PUMAs are definable in simple terms—such as those in the figures that are composed by combining entire counties—but some are based on several geographic components at different levels of Census geography.

The Missouri SDC has developed an online tool, called MABLE/Geocorr, that enables data users to enter the geography that they are interested in to identify PUMA codes and equivalent geographic areas (http://mcdc2.missouri.edu/websas/geocorr2k.html).

The Census Bureau and the SDCs are currently working on revised PUMA geography based on 2010 Census data. Table 4 is a summary of the new criteria for PUMA definition, compared with the previous criteria (U.S. Census Bureau Geography Division 2011). In addition to these criteria, the Census Bureau has provided the SDCs with the

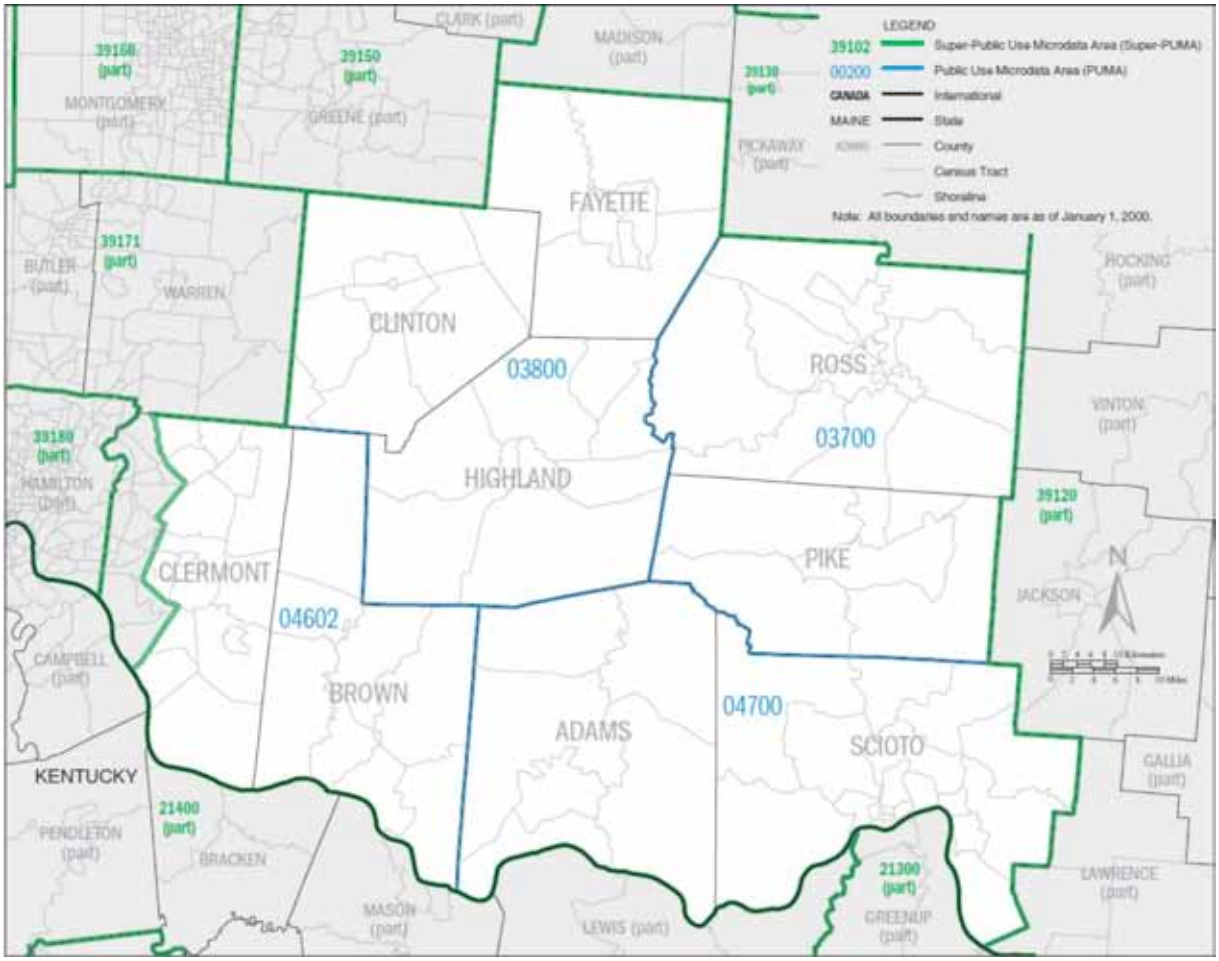


FIGURE 2 An example of a Census Public Use Microdata area map. *Source*: http://www.census.gov/geo/www/maps/puma5pct.htm.

following guidelines (U.S. Census Bureau Geography Division, 2011):

- Wherever possible, each PUMA should comprise an area that is either entirely inside or entirely outside metropolitan or micropolitan statistical areas.
- The Census Bureau recommends that 2010 place definitions, 2000 urban/rural definitions, and local knowledge inform PUMA delineations.
- The number of standard PUMAs should be maximized, and PUMAs should not contain more than 200,000 persons, wherever possible, unless the PUMA is defined for an area in which population decline is anticipated.
- PUMAs should avoid unnecessarily splitting Native American reservations and/or off-reservation trust lands, and separating Native American populations, particularly if large numbers of Native Americans are included within all parts of the split areas.
- To improve the utility and meaningfulness of the PUMAs and the PUMS data, SDCs have the option to provide names for PUMA areas. In addition, the Census Bureau will attempt to maintain previous PUMA numbers for PUMAs whose geographic boundaries are the same or similar to previous boundaries.

The most significant change for the 2010 PUMA delineation is the elimination of Census Place and MCD as building blocks for PUMAs. The Census Bureau has determined that because changes in PUMA building block geography represent a disclosure risk for PUMS publication, and a majority (60%) of all place-based 2000 PUMAs had annexations/de-annexations from 2000 to 2010, it will not permit the use of incorporated places as building blocks for 2010 PUMAs. The Census Bureau reasoned that Census tracts provide much more stable boundaries, and may be aggregated to approximate the extent of other types of geographic entities. It concluded that because PUMS data are subject to noise (i.e., small amounts of variation) and data swapping, they are less susceptible to the small differences between a census tract boundary and an incorporated place or MCD boundary, and therefore these differences are not likely to have a significant impact on the representation of the PUMS data for an incorporated place or MCD (U.S. Census Bureau Geography Division 2011).

The fields and content of the PUMEQ5 file is described below.

FIELD	Description
A	Summary Level Code
B	FIPS State Code
C	SuperPUMA Code
D	PUMA Code
E	FIPS County Code
F	FIPS County Subdivision Code
G	FIPS Place Code
H	Central city indicator
	0 = not in central city
	1 = in central city
I	Census Tract Code
J	Metropolitan Statistical Area / Consolidated Statistical Area Code
K	Primary Metropolitan Statistical Area Code
L	Census 2000 100% population count
M	Area name

A	B	C	D	E	F	G	H	I	J	K	L	M
780	39	39172	03700								101040	PUMA 03700
781	39	39172	03700	131				9999			27695	Pike County
781	39	39172	03700	141				9999			73345	Ross County
780	39	39172	03800								109851	PUMA 03800
781	39	39172	03800	027				9999			40543	Clinton County
781	39	39172	03800	047				9999			28433	Fayette County
781	39	39172	03800	071				9999			40875	Highland County
780	39	39172	04602								115798	PUMA 04602
781	39	39172	04602	015				1642	1640		42285	Brown County
781	39	39172	04602	025				1642	1640		73513	Clermont County (part)
782	39	39172	04602	025	04157			1642	1640		17503	Batavia township
782	39	39172	04602	025	28224			1642	1640		4348	Franklin township
782	39	39172	04602	025	31010			1642	1640		13663	Goshen township
782	39	39172	04602	025	37716			1642	1640		2576	Jackson township
782	39	39172	04602	025	49322			1642	1640		55	Miami township (part)
782	39	39172	04602	025	51335			1642	1640		8236	Monroe township
782	39	39172	04602	025	74825			1642	1640		5316	Stonelick township
782	39	39172	04602	025	75155			1642	1640		5935	Tate township
782	39	39172	04602	025	81130			1642	1640		2361	Washington township
782	39	39172	04602	025	82110			1642	1640		5025	Wayne township
782	39	39172	04602	025	86302			1642	1640		5005	Williamsburg township
780	39	39172	04700								106526	PUMA 04700
781	39	39172	04700	001				9999			27330	Adams County
781	39	39172	04700	145				9999			79196	Scioto County

FIGURE 3 An example of a Census Public Use Microdata area equivalency file. *Source*: http://www.census.gov/main/www/pums.html.

Because the PUMA delineation criteria and population patterns have changed over the past decade, it is expected that the 2010 PUMAs will be significantly different from the 2000 PUMAs (L. Gaines, personal communication, July 2011).

Table 5 shows the schedule for 2010 PUMA delineation. In the autumn of 2011, the Census Bureau will request that the SDCs use specialized software developed by Caliper Corporation called MAF/TIGER Partnership Software to propose PUMA boundaries by early January 2012. The Census Bureau strongly recommends that the SDCs gather local input for PUMA delineations in all areas (U.S. Census Bureau Geography Division 2011).

Consequently, there is an immediate opportunity and need for state DOTs and MPOs to influence PUMA geography.

TABLE 4
CENSUS 2000 PUMA CRITERIA AND NEWLY PUBLISHED PUMA CRITERIA FOR THE 2010 CENSUS

Criterion	Census 2000 Criteria	2010 Census Criteria
PUMA/state relationship	PUMAs may not cross state boundaries	PUMAs may not cross state boundaries
Levels of PUMA geography	Two levels of PUMA geography corresponded to the two types of PUMS files available (1 percent super-PUMA and 5 percent standard PUMA)	One level of PUMA geography corresponds to the one type of decennial and ACS PUMS available
PUMA minimum population threshold	100,000 persons	100,000 persons both at the time of delineation and expected throughout the decade for the publication of ACS PUMS data
PUMA geographic "building block" entities	Counties or equivalent entities, census tracts, incorporated places with populations of 100,000 persons or greater, and county subdivisions (minor civil divisions) in the six New England states	Counties or equivalent entities and census tracts only
PUMA–county part minimum threshold	Not a requirement	Each single PUMA-county part meets a minimum population threshold of 2,400 persons (i.e., the Census 2010 minimum population threshold for a census tract of 1,200 × 2)
PUMA contiguity	Not a requirement as incorporated places can be non-contiguous	A PUMA may be noncontiguous only if a county or a census tract used as a building block for the PUMA is noncontiguous

Source: U.S. Census Bureau Geography Division (2011).

TABLE 5
2010 PUMA DELINEATION SCHEDULE

Date	Activity
July 5, 2011	Final PUMA delineation criteria and guidelines distributed
September 2011	Materials sent to state data centers (SDCs) for PUMA delineation; PUMA delineation software (MTPS) WEBINAR training begins
Late December 2011– Early January 2012	Return deadline for submissions from SDCs
Fall 2011–Spring 2012	Review of PUMA submissions at the Census Bureau and insertion into TIGER database
Spring–Summer 2012	Creation of geographic products containing 2010 PUMAs for use in Decennial Public Use Microdata Sample (PUMS) and American Community Survey (ACS) products
Spring 2012	TIGER/Line® Shapefiles released for 2010 PUMAs
To Be Determined	Decennial (2010) PUMS files, ACS PUMS (1-year, 3-year, 5-year), and ACS estimates (1-year, 3-year, 5-year) released

Source: U.S. Census Bureau Geography Division (2011).

Some transportation planners have found that PUMA delineations provide convenient sets of large subregional districts for data reporting and summary, even if their technical analyses do not necessarily rely on the PUMS data (Metropolitan Transportation Commission Planning Section 2003, Heither 2011). For 2005 and beyond, PUMAs are a standard Census reporting geography, so data users will be able to obtain PUMA-level tabulations directly from American FactFinder. Therefore, Census data users can benefit from meaningfully delineated PUMAs even if they do not currently anticipate future PUMS data usage.

Place of Work Information in Public Use Microdata Sample

POW-PUMAs are different Census geographic delineations that are built from the PUMAs and super-PUMAs. POW-PUMAs are designed to allow for reporting of journey-to-work patterns—PUMS data records contain flows from home origin PUMAs to workplace destination POW-PUMAs, based on Census respondents' reported workplace for the previous week. The recorded POW-PUMA usually corresponds to a work commute destination, but for respondents who travel for work, the POW-PUMA may represent a temporary workplace, which may be quite distant from the respondents' home locations.

POW-PUMAs are most often county based, but can also be defined to the place level or MCD (in the six New England states) (Murakami 2009). An equivalency of PUMA to POW-PUMA can be found in Appendix N of the Census 2000 technical documentation: http://www.census.gov/prod/cen2000/doc/pums.pdf (U.S. Census Bureau 2009a).

Migration Information in Public Use Microdata Sample

MIG-PUMAs are similar to POW-PUMAs, but they relate to place of residence information. MIG-PUMAs are based on counties or (in the six New England states) MCDs, but are not place based (Murakami 2009). MIG-PUMAs are used in the PUMS files to report where Census respondents lived 1 year before they participated in the Census data collection. Understanding the characteristics of recent movers can help

transportation planners identify regional population trends as part of many broader applications. A PUMA to MIG-PUMA equivalency can also be found in Appendix N of the Census 2000 technical documentation: http://www.census.gov/prod/cen2000/doc/pums.pdf (U.S. Census Bureau 2009a).

Workplace Allocation for Public Use Microdata Sample

Once the 2010 PUMAs are established as described previously, the Census Bureau will delineate the POW-PUMAs and MIG-PUMAs. Both of these are expected to be county based, consisting of either a single county-based PUMA or a combination of adjacent tract-based PUMAs that together comprise one or more counties (http://www.census.gov/geo/puma/puma_tutorial.txt) (B. McKenzie, personal communication, July 2011).Thus, PUMS files' PUMA-to-POW-PUMA commuter flow data are usable only at the very large district level. AASHTO's CTPP files are a better source of detailed commuter flows from the Census Journey-to-Work data.

The Census Bureau is able to geocode about three-quarters of the workplace locations provided in the ACS to a specific (census block level) location. The remaining workplace locations are currently assigned to Census Places through a standard allocation procedure. In the future, beginning with the 2012 ACS, the allocation procedure will be extended so that all workplaces will be assigned to the block level. This extended allocation process will also be applied retrospectively to the 2006–2010 ACS-based CTPP data sets, greatly enhancing the usefulness of that data (B. McKenzie, personal communication, July 2011).

Extended workplace allocation is likely to be of less importance for the PUMS than for CTPP because of the large POW-PUMA sizes. A high percentage of the workplace records that are not geocodable to the block level can be geocoded to at least the county level and thus can be assigned to most POW-PUMAs, except those that are census place based. The standard procedures that allocate workplaces to the place level are likely to be adequate for the less detailed POW-PUMA geography. Even so, the accuracy and reasonableness of the allocation processes (both standard and extended) will affect the PUMS data accuracy. The PUMS files do not include a workplace allocation flag, so it is not possible for PUMS users to determine which records have allocated workplace information (B. McKenzie, personal communication, July 2011). Once extended workplace allocation is established as a regular event in ACS production, the Census Bureau will have the option of redefining POW-PUMAs to be the same as standard PUMAs.

OBTAINING PUBLIC USE MICRODATA SAMPLE DATA

As noted previously, the Census Bureau provides the most current PUMS data files on its website (see Figure 4). These data are provided in ASCII text file format as comma-separated values (CSV) files, and in UNIX and PC format SAS data sets. Most statistical software packages such as SAS, SPSS, and Stata can import files of these types. However, as the computing environment evolves, it is becoming easier to obtain PUMS data without relying on these packages, which require a fair amount of training and familiarization and require annual licensing fees. There are two no-cost alternatives: using open-source software such as R to process the data; and relying on online data access systems such as the Census Bureau's DataFerrett and the University of Minnesota's Integrated Public Use Microdata Series (IPUMS), which provide point-and-click table generation without requiring any special programming or software knowledge. These options are discussed here.

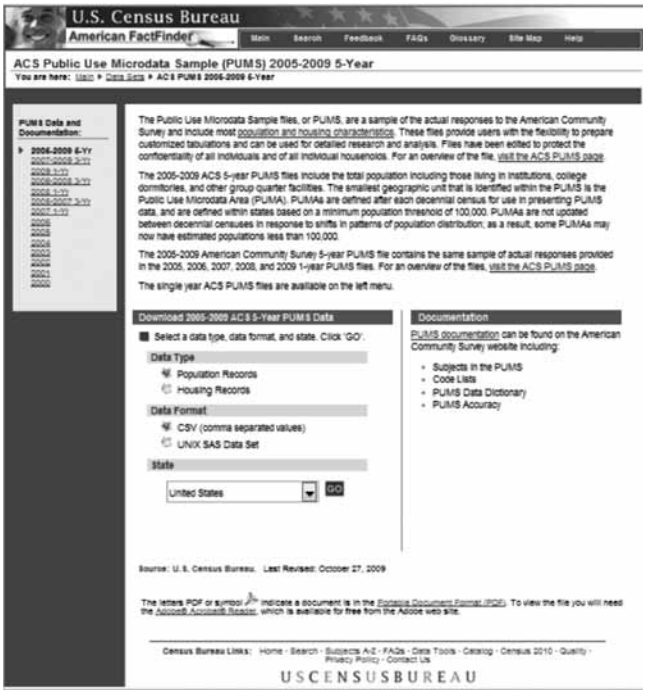


FIGURE 4 Census Bureau website access to PUMS data. *Source*: http://factfinder.census.gov/home/en/acs_pums_2009_5yr.html.

It is more difficult to read the data into spreadsheet and database software packages because of the large number of records in PUMS files. Most PUMS data users obtain the data files directly from the Census Bureau using a statistical software package or an online access system (U.S. Census Bureau 2009a).

The Census Bureau also provides an online utility called DataFerrett (Figure 5) to allow data users to perform tabulations of PUMS data without needing statistical software packages (http://www.census.gov/acs/www/data_documentation/data_ferrett_for_pums/) (U.S. Census Bureau 2009a).

As Figures 6 and 7 show, DataFerrett users can specify PUMS-based tables and variables using point-and-click selections and drop-down lists. Once the user completes the selection, DataFerrett will provide tabulations such as the one shown in Figure 8.

The Census Bureau provides extensive technical documentation for the PUMS data on its website (see Figure 9). This documentation includes the following:

- Subject lists for the various PUMS files,
- Code lists for PUMS variables,
- Top-coded and bottom-coded values for PUMS variables,
- PUMS Data Dictionaries,
- Accuracy of the PUMS memos describing technical aspects of using the PUMS data, and
- PUMS estimates for user verification, including variable tabulations with which users can compare results to ensure that the data are being read successfully.

FIGURE 5 DataFerrett: Census Bureau online tool to access to PUMS data. *Source*: http://www.census.gov/acs/www/data_documentation/data_ferrett_for_pums/.

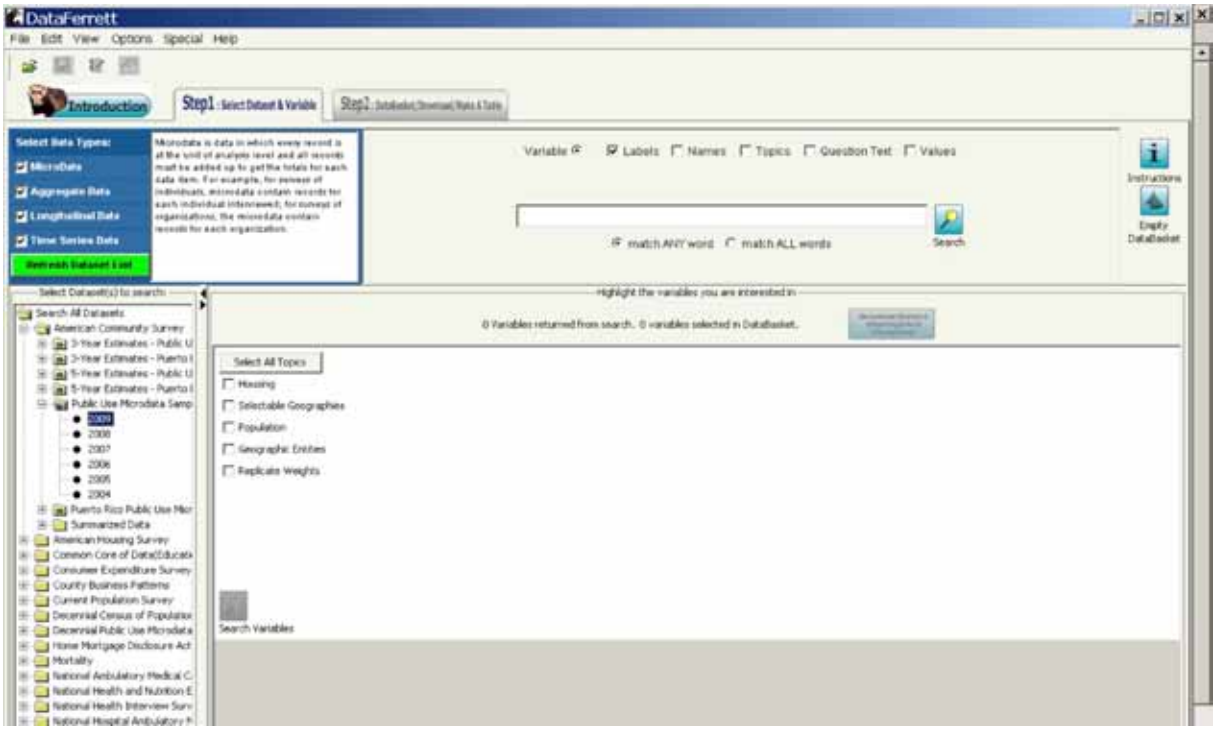


FIGURE 6 DataFerrett PUMS table definition. *Source*: http://www.census.gov/acs/www/data_documentation/data_ferrett_for_pums/.

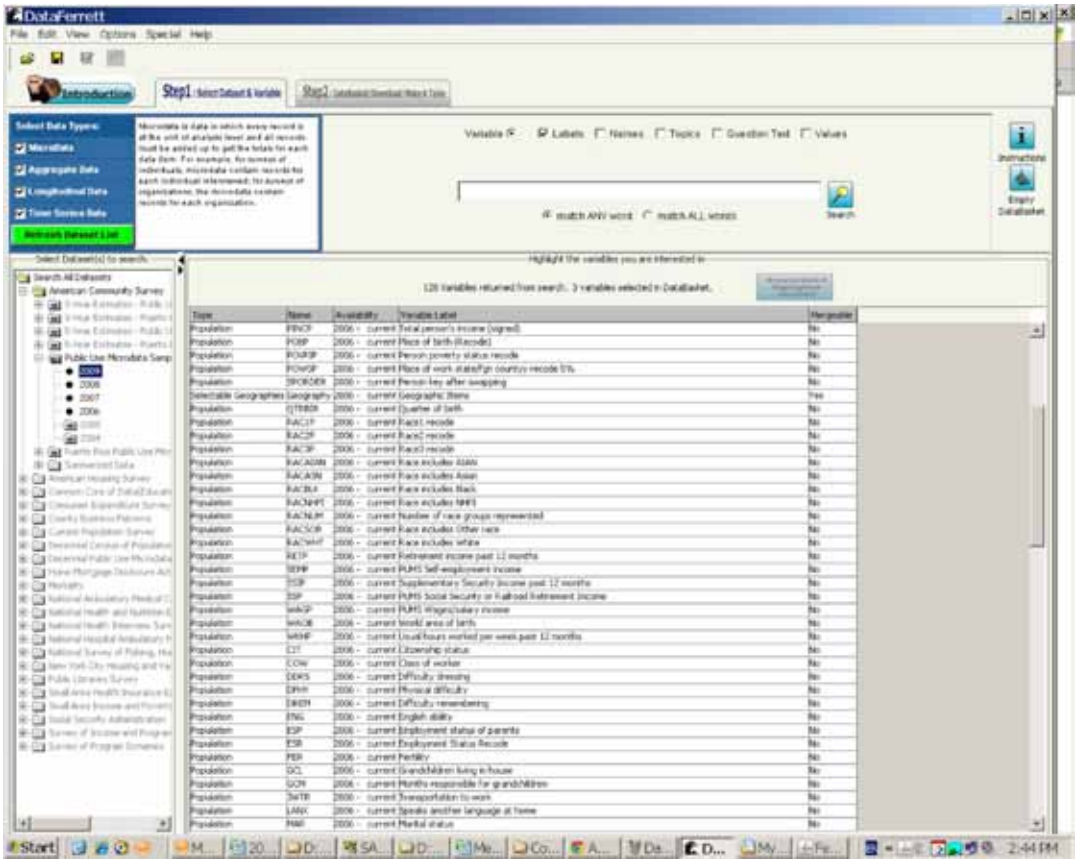


FIGURE 7 DataFerrett PUMS variable definition. *Source*: http://www.census.gov/acs/www/data_documentation/data_ferrett_for_pums/.

	State DOTs	Larger MPOs	Smaller MPOs	Academic Researchers
Travel Demand Modeling	●●●	●●●	●●	●
Travel Surveys	●●●	●●●	●	●
Synthetic Population Microsimulation	●●	●●●	●	●●●
Custom Tabulations	●	●●	●	●●●

●	Minor usage of PUMS
●●	Moderate usage of PUMS
●●●	Significant usage of PUMS

FIGURE 8 PUMS data uses by agency type. *Source*: Synthesis project web-based survey and follow-up research (2011).

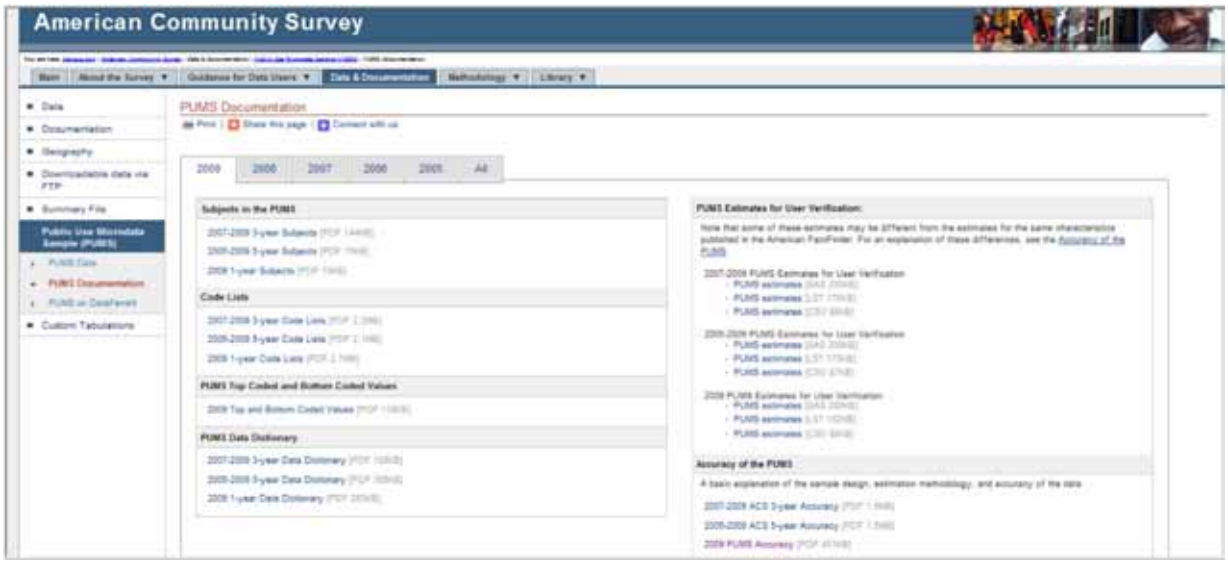
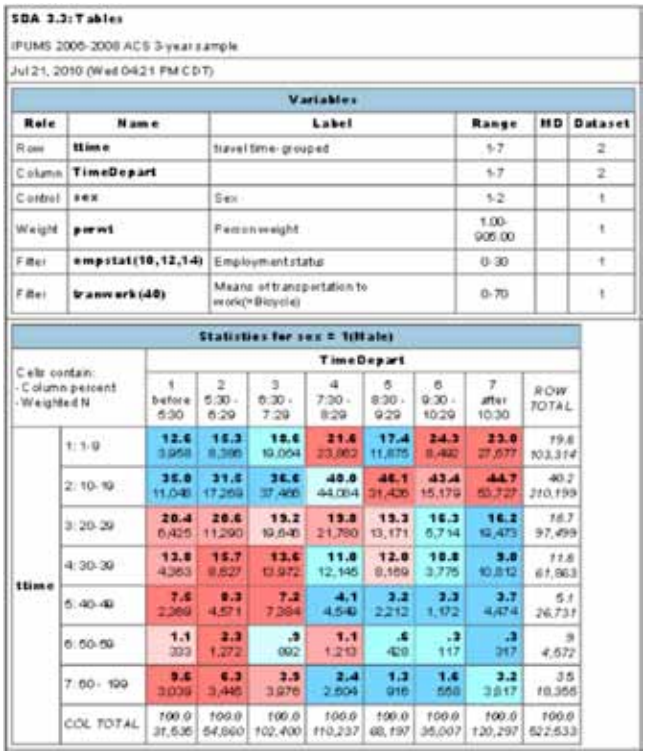


FIGURE 9 Census Bureau website-based PUMS data documentation. *Source*: http://www.census.gov/acs/www/data_documentation/pums_documentation/.

In addition to the Census website sources, affiliate organizations of the Census SDCs provide access to PUMS data and to value-added analysis tools. One of these organizations, the Minnesota Population Center, has developed IPUMS, which combines surviving Decennial Census data from as far back as 1850 and all available ACS PUMS databases (http://usa.ipums.org/usa/). IPUMS reconciles, to the extent possible, differing coding schemes and record layouts from the different PUMS data sources, enabling data users to compare and analyze different years' data more easily.

IPUMS users may obtain data through an online data extraction system that prompts them to select the variables and geography of interest, and then to obtain custom tables based on these variables (see Figure 10). The data are generated on the IPUMS server and can be downloaded for further analysis. IPUMS provides accompanying SAS, SPSS, and Stata syntax files that can be used to easily convert the ASCII files that are generated. IPUMS is freely available without restriction, but users must register before obtaining data. Alternatively, the IPUMS Online Data Analysis System allows users to analyze all IPUMS-USA microdata files online. IPUMS includes an online code book for Census PUMS data and the ability to collapse variables to user-specified categories, which users can save for future use. Figures 11 and 12 show IPUMS output generated by Elaine Murakami in 2010 to help analyze bicycle commuting behavior.

FIGURE 10 IPUMS website access to PUMS data. *Source*: http://usa.ipums.org/usa/.



SDA 3.3: Tables

IPUMS 2006-2008 ACS 3-year sample

Jul 21, 2010 (Wed 04:21 PM CDT)

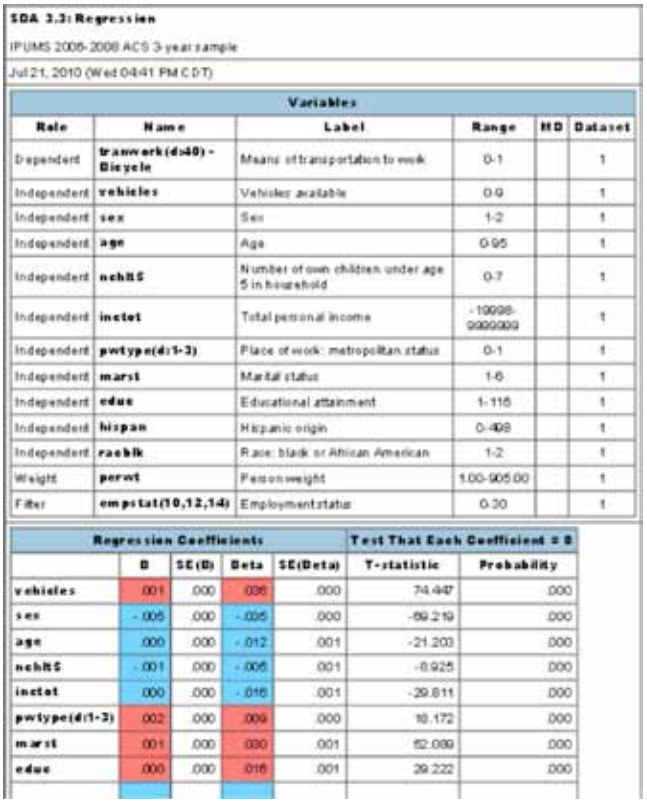
Variables

Role	Name	Label	Range	MD	Dataset
Row	**ttime**	travel time- grouped	1-7		2
Column	**TimeDepart**		1-7		2
Control	**sex**	Sex	1-2		1
Weight	**perwt**	Person weight	1.00-905.00		1
Filter	**empstat(10,12,14)**	Employment status	0-30		1
Filter	**tranwork(40)**	Means of transportation to work(=Bicycle)	0-70		1

Statistics for sex = 1(Male)

Cells contain: - Column percent - Weighted N		TimeDepart							
		1 before 5:30	2 5:30 - 6:29	3 6:30 - 7:29	4 7:30 - 8:29	5 8:30 - 9:29	6 9:30 - 10:29	7 after 10:30	ROW TOTAL
ttime	1: 1-9	**12.6** 3,958	**15.3** 8,396	**18.6** 19,064	**21.6** 23,862	**17.4** 11,875	**24.3** 8,492	**23.0** 27,677	*19.6* *103,314*
	2: 10-19	**35.0** 11,048	**31.5** 17,269	**36.6** 37,466	**40.0** 44,084	**46.1** 31,436	**43.4** 15,179	**44.7** 53,727	*40.2* *210,199*
	3: 20-29	**20.4** 6,425	**20.6** 11,290	**19.2** 19,646	**13.8** 21,780	**13.3** 13,171	**16.3** 5,714	**16.2** 19,473	*16.7* *97,499*
	4: 30-39	**13.8** 4,363	**15.7** 8,627	**13.6** 13,972	**11.8** 12,145	**12.0** 8,169	**10.8** 3,775	**9.8** 10,812	*11.8* *61,863*
	5: 40-49	**7.6** 2,389	**9.3** 4,571	**7.2** 7,384	**4.1** 4,540	**2.2** 2,212	**3.3** 1,172	**3.7** 4,474	*5.1* *26,731*
	6: 50-59	**1.1** 333	**2.3** 1,272	**.9** 892	**1.1** 1,213	**.6** 429	**.3** 117	**.3** 317	*.9* *4,572*
	7: 60 - 199	**9.6** 3,039	**6.3** 3,445	**3.9** 3,976	**2.4** 2,604	**1.3** 916	**1.6** 558	**3.2** 3,817	*3.5* *18,355*
	COL TOTAL	*100.0* *31,535*	*100.0* *54,860*	*100.0* *102,400*	*100.0* *110,237*	*100.0* *68,197*	*100.0* *35,007*	*100.0* *120,297*	*100.0* *522,533*

FIGURE 11 Example of IPUMS tabular output. *Source*: http://usa.ipums.org/usa/.



SDA 3.3: Regression

IPUMS 2006-2008 ACS 3-year sample

Jul 21, 2010 (Wed 04:41 PM CDT)

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	**tranwork(d=40) - Bicycle**	Means of transportation to work	0-1		1
Independent	**vehicles**	Vehicles available	0-9		1
Independent	**sex**	Sex	1-2		1
Independent	**age**	Age	0-95		1
Independent	**nchlt5**	Number of own children under age 5 in household	0-7		1
Independent	**inctot**	Total personal income	-19998-9999999		1
Independent	**pwtype(d=1-3)**	Place of work: metropolitan status	0-1		1
Independent	**marst**	Marital status	1-6		1
Independent	**educ**	Educational attainment	1-116		1
Independent	**hispan**	Hispanic origin	0-498		1
Independent	**raceblk**	Race: black or African American	1-2		1
Weight	**perwt**	Person weight	1.00-905.00		1
Filter	**empstat(10,12,14)**	Employment status	0-30		1

Regression Coefficients				Test That Each Coefficient = 0	
	B	SE(B)	Beta	SE(Beta)	T-statistic	Probability
vehicles	.001	.000	.036	.000	74.447	.000
sex	-.005	.000	-.035	.000	-69.219	.000
age	.000	.000	-.012	.001	-21.203	.000
nchlt5	-.001	.000	-.005	.001	-8.925	.000
inctot	.000	.000	-.016	.001	-29.811	.000
pwtype(d=1-3)	.002	.000	.009	.000	18.172	.000
marst	.001	.000	.030	.001	52.089	.000
educ	.000	.000	.016	.001	29.222	.000

FIGURE 12 Example of IPUMS regression modeling output. *Source*: http://usa.ipums.org/usa/.

In one side-by-side comparison, a transportation planner who was new to the use of PUMS data found that obtaining data through IPUMS was preferable to obtaining the data directly from the Census Bureau (Azimi 2005). In April 2009, FHWA offered a training webinar on IPUMS. The webinar materials are available on the FHWA/AASHTO CTPP website, http://ctpp.transportation.org/Pages/webinardirectory.aspx.

CHAPTER THREE

SURVEY OF USAGE OF THE PUBLIC USE MICRODATA SAMPLE DATA BY TRANSPORTATION PLANNERS

WEB-BASED SURVEY SCAN OF TRANSPORTATION PLANNERS

Although the main goal of the synthesis was to identify recent and ongoing usage of the PUMS data, it also sought to determine overall usage levels of the data generally. To better understand how many different transportation planning agencies use the PUMS data, and for what purposes, a web-based scan of state DOTs and MPOs was conducted. As Table 6 shows, all the state DOTs and larger MPOs were invited to participate, as it was believed that these agencies would be the most likely to be active PUMS data users and therefore would be more likely to be able to report PUMS activities. The sample of smaller agencies was randomly selected from the list of current MPOs on the U.S. DOT Transportation Planning Capacity Building website (http://www.planning.dot.gov/default.asp). The specific contact list for the state DOTs was compiled by NCHRP staff from the Standing Committee on Planning organization list. Specific MPO contacts were identified from staff directories found on agency websites. Where possible, MPO staff members with Census data responsibilities were selected. If no Census specialist could be defined, a planning director or similar official was selected as the recipient of the emailed survey invitation.

TABLE 6
WEB-BASED SURVEY SAMPLE AND RESPONSE

Agency Type	Total Agencies	Sample Agencies	Completed Surveys	Response Rate
State DOTs	52	52	37	71%
Large MPOs	42	42	23	55%
Small/Medium MPOs	339	75	25	33%
Total	433	169	85	50%

Note: Large MPOs were defined as those that support populations of more than 1 million. No distinction between small MPOs (representing populations of 200,000 or less) and medium MPOs was made for these summaries.
Source: Synthesis project web-based survey (2011).

The web-based survey was limited in a few important respects:

- Identification of the proper survey respondents within agencies was extremely difficult because of the agency sizes and decentralized functions. Many respondents noted that others within their agency might have relevant Census PUMS usage, but that they were unsure who they were. It is very likely that response rates and accuracy were diminished because the web survey invitations did not reach all of the relevant people within the agencies.
- The nature of the web survey recruitment is likely to have introduced bias into the survey results. As PUMS data usage is not common, many potential web survey respondents were being asked to participate in a survey regarding a topic in which they had little interest or knowledge. Consequently, those with PUMS data usage experience were more likely to participate.

Therefore, the web-based scan is not to be viewed as a scientific survey. Rather, this data collection effort sought to identify general comparative trends in Census PUMS data usage and to organize those data uses.

The response rate from state DOTs and larger MPOs was significantly higher than that for smaller MPOs, even before follow-up activities. This is likely the result of a number of factors, including the larger agencies' greater familiarity with Census data products and NCHRP research efforts, as well as the higher likelihood that the sample identified relevant respondents at the larger agencies. Almost all the survey respondents indicated that several staff agency members used Census data products, but the majority of respondents claimed to be aware of the Census-based analyses that were taking place at the agency even if they were not directly involved in the analyses.

FAMILIARITY AND USAGE OF PUBLIC USE MICRODATA SAMPLE DATA

Table 7 summarizes the usage of PUMS data by the responding agencies. Slightly more than a third of the responding state DOTs are regular or occasional users of PUMS data. About two-thirds of the large MPOs that responded use PUMS, and most of these do so regularly. However, fewer than 15% of small and medium MPOs that responded to the survey scan use these data. Ten of the 37 responding state DOTs, five of 25 smaller MPOs, and four of 23 large MPOs were not familiar with the PUMS data source.

TABLE 7
USAGE OF CENSUS PUMS DATA BY TRANSPORTATION PLANNING AGENCIES

Agency Usage of PUMS Data	State DOTs	Large MPOs	Small/Medium MPOs
Regular User of This Data Source	2	12	1
Occasional User of This Data Source	11	3	2
Familiar with This Data Source but Do Not Use It	14	4	17
Not Familiar with This Data Source	10	4	5
Total	37	23	25

Source: Synthesis project web-based survey (2011).

To put this level of usage and familiarity into context, respondents were asked about their usage of and familiarity with several other Census Bureau and related data sets. As Table 8 shows, PUMS usage is relatively low compared with other Census data sets and other similar data sources. This is particularly true for smaller to medium-sized MPOs and state DOTS.

TABLE 8
RESPONDENTS INDICATING REGULAR OR OCCASIONAL USAGE OF DIFFERENT DATA SETS BY TRANSPORTATION PLANNING AGENCIES

Data Source	State DOTs ($N = 37$)	Large MPOs ($N = 23$)	Small/Medium MPOs ($N = 25$)
Decennial Census Tabulations	35	23	23
American Community Survey (ACS) Tabulations	29	19	20
Census Transportation Planning Products (CTPP)	26	20	16
Census Annual Population Estimates	24	20	12
State or Local Household Travel Survey(s)	18	20	12
FHWA National Household Travel Survey (NHTS)	20	19	9
Bureau of Labor Statistics Employment Databases	15	19	7
Census Public Use Microdata Sample (PUMS)	13	15	3
Census Economic Surveys	12	11	4
Census Longitudinal Employer-Household Dynamics (LEHD)	7	12	1

Source: Synthesis project web-based survey (2011).

These results indicate that PUMS data usage is not widespread among transportation planners. Results on the reasons for this usage level are discussed here. However, despite the limited level of usage, the agencies that do use the data find them to be quite valuable. As Table 9 shows, all but one respondent at agencies that use the PUMS data indicated that the PUMS data played a very important or somewhat important part in helping their agency fulfill its mission.

TABLE 9
IMPORTANCE OF THE PUMS DATA FOR TRANSPORTATION PLANNING AGENCIES THAT USE THEM

Agency Rating of PUMS Data	State DOTs	Large MPOs	Small/Medium MPOs
Very Important/ Central to Agency's Mission	4	12	2
Somewhat Important	9	2	1
Useful, but Not Too Important	0	1	0
Not Useful	0	0	0

Source: Synthesis project web-based survey (2011).

Respondents were asked to broadly classify their uses of the PUMS data. As shown in Table 10, most agencies use the PUMS to support travel demand modeling. More than half of state DOT PUMS data users also use PUMS to support travel survey efforts, and a majority of large MPOs that use PUMS data use them to support synthetic population microsimulation.

TABLE 10
PUMS DATA USES OF TRANSPORTATION PLANNING AGENCIES

Uses of PUMS Data	State DOTs ($N = 13$)	Large MPOs ($N = 15$)	Small/Medium MPOs ($N = 3$)
Travel Demand Modeling Components	11	10	3
Weighting and Expansion of Travel Surveys	7	3	1
Synthetic Population Microsimulation	4	11	1
Custom Census Data Cross-Tabulations	2	6	0

Source: Synthesis project web-based survey (2011).

In addition to the classification of the data uses, respondents listed the following individual uses of the Census PUMS data:

- Analyze determinants of income and inequality,
- Commutation data by POW-PUMA by mode,
- Comparison with household travel survey estimates,
- Limited English proficiency analysis,
- Predicting housing type choice (e.g., renters, owners),
- Socioeconomic/land use modeling,
- Development of submodels (e.g., income, car ownership, size, worker, age of head of household, children),
- Environmental justice analyses,
- Estimates of commuters to/from MPO modeled area from external commuting shed,
- Household characteristics (e.g., age of head of household, number of workers, number of persons, share of households by income),
- Commuter characteristics,
- Departure time for primary work location,
- Estimates of "out of town" and self-employed workers, and
- Statewide artificial population synthesizer.

These data uses are described in chapter four.

PUMS data users tend to obtain their data directly from the Census Bureau, rather than from other third-party sources of the PUMS data such as IPUMS (Table 11). Most of the PUMS data users who were interviewed for this project have expertise with statistical software packages, such as SAS and SPSS. Many indicated a strong preference for the direct provision of data, rather than obtaining data through programs such as DataFerrett or IPUMS. PUMS data usage appears to be largely limited to the minority of agencies that have developed internal experts in the use of the PUMS products.

TABLE 11
MEANS OF ACCESSING PUMS DATA BY TRANSPORTATION PLANNING AGENCIES

Agency Access of PUMS Data	State DOTs	Large MPOs	Small/Medium MPOs
Directly from the Census Bureau website or data center?	11	14	3
From a university or company that disseminates the data in its original format?	1	1	0
Grand total	12	15	3

Source: Synthesis project web-based survey (2011).

DATA USERS' ATTITUDES TOWARD PUBLIC USE MICRODATA SAMPLE DATA

As shown in Tables 12–14, responding PUMS data users, for the most part, regard accessing, manipulating, analyzing, and obtaining documentation on PUMS as straightforward. The Census Bureau provides the PUMS data in SAS format and in text file format.

The data users who were interviewed separately from the web-based survey also rated the ease of use of PUMS highly, but many of these respondents were already familiar with the PUMS data.

TABLE 12
DATA USERS' RATING OF EASE OF ACCESSING PUMS DATA

Ease of Accessing the PUMS Data	State DOTs	Large MPOs	Small/Medium MPOs
Excellent	3	3	0
Good	6	10	0
Fair	2	2	2
Poor	0	0	0
No Opinion/Not Sure	2	0	1
Total	13	15	3

Source: Synthesis project web-based survey (2011).

TABLE 13
DATA USERS' RATING OF EASE OF MANIPULATING AND ANALYZING THE PUMS DATA

Ease of Manipulating and Analyzing the PUMS Data	State DOTs	Large MPOs	Small/Medium MPOs
Excellent	1	2	0
Good	9	10	1
Fair	1	3	1
Poor	0	0	0
No Opinion/Not Sure	2	0	1
Total	13	15	3

Source: Synthesis project web-based survey (2011).

TABLE 14
DATA USERS' RATING OF AVAILABILITY AND QUALITY OF THE DOCUMENTATION FOR THE PUMS DATA

Availability and Quality of the Documentation for the PUMS Data	State DOTs	Large MPOs	Small/Medium MPOs
Excellent	1	3	1
Good	10	10	0
Fair	0	2	0
Poor	1	0	1
No Opinion/Not Sure	1	0	1
Total	13	15	3

Source: Synthesis project web-based survey (2011).

The PUMS data users in the web-based survey were asked to provide suggested improvements (up to four per respondent) for the PUMS data. They had a wide range of suggestions about several different aspects of the PUMS database and data delivery, including the following:

- Additional PUMS Data Items
  - "Use of NAICS codes (at 2 or 3 digit level) in defining employment."
  - "It's kind of a pain to do MSA-level tabulations, might be nice if there were an MSA field."
  - "Enhance the richness of the data."
- PUMA Geography
  - "Finer resolution in PUMA geography (i.e., smaller PUMAs)."
  - "More refined, smaller geography."
  - "PUMAs are too large."
  - "PUMA size/detail."
  - "PUMA boundaries not to cross MPO boundaries."
  - "Multi-county PUMA definitions are not convenient for many analyses."
  - "MPO input into defining PUMA boundaries."
  - "Historical consistency in PUMA definitions is important."
- Place-of-Work PUMAs
  - "Place-of-Work geocoded to Place-of-Residence PUMA."

  - "Place-of-Work PUMA should be the same as Place-of-Residence PUMAs."
  - "The POW-PUMA is by jurisdiction and not at the detail of resident PUMA."
- Data Access and Presentation
  - "Availability of better analysis and tabulation tools."
  - "It would be nice if it were possible to specify which fields you want to download."
  - "Look to IPUMS as to what you should be doing."
  - "Keep it current and up to date more often."
  - "Presentation ready output."
  - "Visualization tools."
  - "Promote innovative use of the data."
- Concerns about PUMS with Migration to ACS
  - "With ACS the PUMS is annually updated. Are there examples of anyone using the data to track change? Can PUMS data be used in performance measures?"
  - "Guidance for calculating standard errors using single-year versus multi-year ACS PUMS data."
  - "Concerned about sample sizes going forward with ACS."
- Documentation Improvements
  - Keep the PUMS data format/code consistent over the time (they changed from 1990 to 2000).
  - "New releases of PUMS change categories, so we have to verify in the data dictionary that our procedures still work. They should release multiple versions of the data so that they are backwards compatible."
  - Better documentation of use of PUMS data in transportation planning.
  - A crosswalk between the Census 2000 variable names to the ACS variable names.
  - Instructions/studies on appropriate and inappropriate uses of PUMS data.

The data users were then asked for their opinions of different specific aspects of the PUMS data, as shown in Tables 15–18. These ratings categories reflect common criticisms and limitations of the PUMS data found in the literature, but as was the case for the assessments related to the ease of using the PUMS data, the data users generally rate the different aspects of the PUMS data well. This probably indicates an acceptance, or at least an understanding, of the reasons for the apparent weaknesses of the PUMS data by users who are still able to use the data to achieve their specific analyses.

Among the PUMS data users who were interviewed separately from the web-based survey, the most commonly raised limitation of PUMS was the need to have large geographic areas. Several respondents believed that the 100,000 minimum population limits were arbitrarily large, and noted that they would like smaller geographic areas to be reviewed. Nonetheless, most of these data users understood and were sympathetic to the need for data disclosure limits, such as minimum population sizes for PUMAs and the perturbation of PUMS data records.

TABLE 15
DATA USERS' RATING OF PUMS SAMPLE SIZES

PUMS Sample Sizes	State DOTs	Large MPOs	Small/Medium MPOs
Excellent	2	1	0
Good	8	11	0
Fair	1	2	1
Poor	0	1	1
No Opinion/Not Sure	2	0	1
Total	13	15	3

Source: Synthesis project web-based survey (2011).

TABLE 16
DATA USERS' RATING OF PUMS GEOGRAPHIC DEFINITIONS

PUMS Geographic Definitions	State DOTs	Large MPOs	Small/Medium MPOs
Excellent	1	1	0
Good	7	9	2
Fair	4	4	0
Poor	0	1	0
No Opinion/Not Sure	1	0	1
Total	13	15	3

Source: Synthesis project web-based survey (2011).

TABLE 17
DATA USERS' RATING OF PUMS DATA DISCLOSURE AVOIDANCE PROCEDURES

PUMS Data Disclosure Avoidance Procedures	State DOTs	Large MPOs	Small/Medium MPOs
Excellent	2	2	0
Good	8	9	0
Fair	2	2	2
Poor	0	1	0
No Opinion/Not Sure	1	1	1
Total	13	15	3

Source: Synthesis project web-based survey (2011).

TABLE 18
DATA USERS' RATING OF PUMS COMMUTING AND WORKPLACE DATA

PUMS Commuting and Workplace Data	State DOTs	Large MPOs	Small/Medium MPOs
Excellent	1	0	0
Good	6	11	0
Fair	4	3	0
Poor	0	0	0
No Opinion/Not Sure	2	1	3
Total	13	15	3

Source: Synthesis project web-based survey (2011).

Data users also expressed some concern over the need to pool data across years of ACS data collection in order to

obtain large samples. There was a concern that in fast-growing communities, multiyear PUMS data would provide an outdated population snapshot. Data users wondered whether single-year PUMS sample sizes could be made larger.

A third area of concern that data users raised is related to the PUMA definition process. A few data users mentioned that the Census Bureau and the SDCs of adjoining states appeared to have different criteria and philosophies of how best to define PUMA geography, making analyses of multistate populations more difficult.

REASONS FOR NOT USING PUBLIC USE MICRODATA SAMPLE DATA

Nonusers of the PUMS data were asked to indicate whether any of the reasons listed in Table 19 helped to explain why they did not use the PUMS data. The most common reason nonusers gave for not using the PUMS data was their lack of familiarity with the data. Roughly half of nonusers are not completely aware of what the PUMS data are. The lack of technical understanding of PUMS also plays an important role in agencies choosing not to use the data.

TABLE 19
AGENCIES' REASONS FOR NOT USING CENSUS PUMS DATA

Reasons for Not Using PUMS Data	State DOTs	Large MPOs	Small/Medium MPOs
Not completely aware of what the PUMS data are	11	4	9
No need for PUMS given availability of other data sources	12	2	12
Lack technical knowledge and/or time needed to use PUMS	6	2	4
Not satisfied with PUMS geographic area sizes	1	0	3
Software and computing limitations	0	0	0
Not satisfied with PUMS quality and consistency	0	0	0
Not satisfied with PUMS weighting and sampling	0	0	0
Not satisfied with PUMS workplace location / commuting	0	0	0
Other reasons for not using PUMS: Analyses performed by other agencies or consultants	2	1	1
Total agencies in sample not using PUMS	24	8	22

Source: Synthesis project web-based survey, 2011.

One-half of the smaller MPOs and about one-third of the state DOTs indicated that they have not yet needed PUMS data, and some agencies rely on other agencies or consultants to perform PUMS-related analyses. In survey follow-up calls, several non-PUMS users indicated that they expected to work with PUMS in the future as their agencies' data and modeling needs change.

Of the perceived limitations and challenges of using the PUMS data, only the large PUMA sizes had an effect on PUMS usage. None of the nonusers indicated that they were kept from using PUMS by software or computing constraints. Lack of satisfaction with PUMS data quality, sampling, or workplace location data also does not play a role in nonusers' decisions not to use the data. The limited survey results indicate that the lack of familiarity with and knowledge of the PUMS data are the primary reasons why many agencies are not using PUMS.

CONCLUSIONS FROM THE WEB-BASED SURVEY SCAN

The web-based survey scan of state DOT, larger MPO, and smaller MPO data users was supplemented with follow-up interviews of transportation planners, researchers, and consultants. In addition, a literature search identified several uses of the PUMS data to support transportation planning. The following conclusions were drawn from the review:

- Among state DOT and MPO analysts, Census PUMS data use is less prevalent than the use of most other Census data products.
- Of the three types of agencies, large MPOs are most likely to use the PUMS data, and smaller MPOs are least likely to use these data.
- Figure 13 summarizes the types of transportation analyses for which data specialists use PUMS.
- In general, agencies that use PUMS data consider these data to be very important or somewhat important to their objectives, and tend to rate the data highly along most quality dimensions.
- Despite the moderately high satisfaction levels of PUMS users, many of the agencies that do not use PUMS data are not aware of what the data are or have not identified a specific need for the data.

The following chapter outlines specific examples of how PUMS data are being used, and describes the benefits and drawbacks of PUMS.

	State DOTs	Larger MPOs	Smaller MPOs	Academic Researchers
Travel Demand Modeling	●●●	●●●	●●	●
Travel Surveys	●●●	●●●	●	●
Synthetic Population Microsimulation	●●	●●●	●	●●●
Custom Tabulations	●	●●	●	●●●

●	Minor usage of PUMS
●●	Moderate usage of PUMS
●●●	Significant usage of PUMS

FIGURE 13 PUMS data uses by agency type. *Source*: Synthesis project web-based survey and follow-up research (2011).

CHAPTER FOUR

APPLICATIONS OF THE PUBLIC USE MICRODATA SAMPLE DATA

COMMON TRANSPORTATION PLANNING USES OF CENSUS DATA

Transportation planners use Census data to support a wide range of functions and planning activities. Understanding the relationships between household and population characteristics information collected by the Census Bureau and transportation system usage data is a key aspect of transportation planning.

The Census Bureau provides several hundred tabulations and cross-tabulations for users on its American FactFinder website and in its downloadable summary files. The table definitions were selected by the Census Bureau with input from data users. The Census Bureau seeks to provide the tabulations that the data user community is most likely to need. (For a list of available tables, see http://www.census.gov/acs/www/data_documentation/2009_5yr_data/.)

In addition, AASHTO's CTPP provides special Census tabulations of particular interest to transportation planners. These tabulations were first developed for the 1970 Decennial Census long form data, and they have been made available for all the Decennial Census long form data collection efforts since then. With the migration to ACS, AASHTO has supported the development of a 3-year (2006–2008) CTPP and will be supporting the production of a 5-year set of tabulations with detailed geographic delineation for the period 2006 to 2010.

The CTPP 2000 and ACS CTPP are divided into three parts:

- Part 1 contains residence end data summarizing worker and household characteristics.
- Part 2 contains place-of-work data summarizing worker characteristics.
- Part 3 contains journey-to-work flow data.

The CTPP tables provide many of the Census tabulations that transportation planners rely on the most. (See http://www.fhwa.dot.gov/ctpp/dataprod.htm for a list of available CTPP tables.) However, owing to the imposition of Census Bureau data disclosure rules, many of the CTPP 2000 tabulation data were suppressed for smaller geographic areas. As the ACS 5-year sample sizes are smaller than the Decennial Census, the data suppression for potential new CTPP products would be even worse. Consequently, efforts are now focused on ways to generate a complete set of data consisting of perturbed values that strive to retain the usability of the CTPP tabulations.

The goal of research project NCHRP 08-79, "Producing Transportation Data Products from the American Community Survey (ACS) that Comply with Disclosure Rules," is to develop a practical approach to perturb ACS data to allow for small-area CTPP tabulations. Preliminary findings of this effort indicate that a workable perturbation approach can be developed. If such an approach were implemented and found to be acceptable by the data user community, transportation planners would be able to obtain a large number of transportation-related tabulations for detailed geography (Krenzke 2010).

COMMON TRANSPORTATION PLANNING USES OF PUBLIC USE MICRODATA SAMPLE FILES

Despite the availability of so many Census and FHWA/AASHTO data resources, some transportation planners and researchers still rely on the disaggregate PUMS data to investigate relationships between Census subjects, either by cross-tabulating the data or by developing statistical relationships between data variables. Based on discussions with PUMS data users and on summaries of research using these data, transportation planners and researchers have found PUMS to be especially useful for the following compilations:

- **Cross-tabulations of variables not readily available from Census or CTPP** – The available Census and CTPP tables often allow transportation planners to easily locate information needed to support planning applications, but some analysis needs require users to combine population characteristics that are not included in the available tabulations. Often, these analyses look at special subpopulations (e.g., members of ethnic groups, people of certain ancestries, group quarters residents, bicycle commuters) that can be separated using the PUMS data.
- **Cross-tabulations of variables in CTPP but with more currency** – Because PUMS data are available on an ongoing basis and the CTPP are available periodically, planners can use the PUMS data to create more

up-to-date CTPP-like data tables, albeit with less precision in the estimates and less geographic detail.

- **Disaggregate analyses** – Planners and modelers frequently require household- or person-level (disaggregate) data to develop models of the interrelationships between household and person characteristics. The microdata represented by the PUMS data allow users to evaluate variable relationships at the housing unit and person levels.
- **Comparisons of different regions** – Because the PUMS data series provides common data sets for all regions of the country, and the Census Bureau provides the same attention to detail in its data collection efforts, the PUMS data are particularly useful for interregional comparisons and national analyses.
- **Comparisons over time** – Because PUMS data sets exist for each of the Decennial Census data collection efforts and each ACS implementation year, the data are commonly used to track changes in housing and person characteristics and changes in the interrelationships between these characteristics over time. Minor changes in the variables and reporting levels make these comparisons more difficult, but planners and researchers are only beginning to explore the utility of annual PUMS data.
- **Validation of other data sources** – PUMS data can be used to independently check calculations and predictions made using other data sources, such as travel surveys, demographic estimates, and modeling results.

Information gathered about recent uses of the data indicates that, PUMS-based cross-tabulations and models are being conducted to address analysis needs across several subject areas. These data analyses frequently support specific issues and studies, such as understanding the commuting characteristics of specific population subgroups or the demographic characteristics of commuters by mode. PUMS analyses are sometimes the main focus of planning analyses, but more frequently they make up only one part of a multistep process.

Described here are several examples of descriptive studies and planning applications for which planners or researchers took advantage of the benefits of the PUMS data listed previously. Also described are some cases in which planners or researchers were limited by PUMS disadvantages. These studies are summarized so that readers may better understand the range of transportation analyses that are available with the PUMS data. The discussions focus on the use of PUMS as opposed to the research findings or the analysis strategies. The original documents from which many of these summaries were derived can provide readers with more background and more information on the substantive analyses and the overall planning projects.

For each of the cited studies, Table 21 lists the reasons that PUMS data were used and key issues identified with the PUMS data usage.

The uses of the PUMS data are organized into four broad data functions:

- User custom tabulations and cross-tabulations of PUMS to support transportation planning decision-making and research;
- Use of PUMS to support travel surveys;
- Use of PUMS to support travel demand modeling efforts; and
- Use of PUMS for population microsimulation.

CUSTOM CROSS-TABULATIONS AND SUMMARIES OF PUBLIC USE MICRODATA DATA

The Census Bureau provides PUMS data files because it recognizes that it could never anticipate all the tabulations and summaries of ACS or Decennial Census data that users might desire. Transportation planners use custom tabulations of PUMS to study a wide range of topics.

Transportation Profiles Using Public Use Microdata

PUMS data can be used to provide an overall picture of a population as it relates to transportation planning issues. Two recent examples are from the Houston-Galveston area and Florida.

Ju (2007, p. 12) conducted an analysis of the demographic changes in the Houston-Galveston area between 2000 and 2005 using Decennial Census and ACS PUMS data. PUMS data were summarized and compared for a large number of variables. The analysis showed several changes in the population characteristics between the two periods. Especially large changes were seen in the disability rate, perhaps because of issues with Census 2000 data or better capture of the disabled in the ACS. In part, these analyses and comparisons were conducted in order to evaluate the ACS PUMS data.

Based on her review, Ju (2007) suggested that it is important that data users understand the differences in the Census data collection efforts to use them properly. When data users look at change over time, they need to be careful about interpretation. Use of standard errors for both time periods will avoid many common pitfalls. The researcher warned that "the data have a margin of error associated, and a lot of the 'apparent' changes seen over that period were not statistically significant when you introduced the margin of error."

Zhou (2004) developed a profile of journey-to-work characteristics for Florida commuters using the year 2000 5% PUMS data, the year 2000 1% PUMS data, and ACS test site PUMS data. The thesis presented detailed comparisons of journey-to-work private vehicle occupancy distribution, travel time distribution, mode choice, and departure time

distribution by different household and individual characteristics. The data analysis showed that the three data files reflected acknowledged demographic trends and captured known changes such as aging of population, smaller household size, and increasing car ownership. The analysis also showed that in most cases, the (then) new ACS PUMS data files approximated the year 2000 PUMS files very well.

More common than general profiles are tabulations of how specific PUMS variables relate to each other. The PUMS data have been used to provide background information for many types of analyses:

- The San Francisco Metropolitan Transportation Commission (MTC) used PUMS data to tabulate the income distribution of trans-bay commuters by means of transportation (bus versus drive alone versus rail) in support of the Bay Bridge Congestion Pricing Study (C. Purvis, personal communication, Jan. 2011).
- MTC used PUMS data to analyze the characteristics (earnings, occupation, industry, sex) of Marin County work-at-home workers (C. Purvis, personal communication, Jan. 2011).
- The Metropolitan Washington Council of Governments (MWCOG) used PUMS data on self-employed workers to improve regional employment estimates and forecasts that were derived from other sources. It also uses PUMS data to analyze labor force participation (R. Griffiths, personal communication, Apr. 2011).
- Several agencies have used PUMS to analyze the relationships between housing unit variables (structure type, age of building, years at the address, tenure) and between these variables and household characteristics (age of householders, household income, vehicle availability) (web-based survey scan 2011).
- Planning agencies and transit agencies have used PUMS to summarize transit commuting levels by the limited English proficiency population (web-based survey scan 2011).

The large number of PUMS variables and the level of variable detail PUMS affords have allowed data users to perform a wide range of detailed cross-tabulations.

FHWA (2002) guidance on environmental justice analysis suggests that analysts take advantage of the PUMS data for job access and reverse commute planning.

> Low-income and welfare-to-work populations are part of a critical transit dependent segment whose needs must be closely evaluated to establish effective transit solutions. Transportation agencies must find ways to bridge the gap between job locations and residences, and to do so effectively, give attention to the training and educational needs and staffing requirements of various industry sectors. Job access studies can take a closer look at the staffing needs—the occupational mix and basic educational attainment requirements of job openings (FHWA 2002).

FHWA (2002) provided two specific examples of how PUMS data have helped in environmental justice analyses:

- For a study in Atlanta, PUMS data helped labor market and transportation analysts distinguish entry-level jobs and other occupations most suitable to persons with limited formal education or training. PUMS reports both occupation and industry employment and, therefore, supports a bridge table or matrix that suggests the types of occupational openings from employment growth in a given industry. Labor agencies and academic institutions have used PUMS to assess the education attainment levels of persons in various industries and occupations (FHWA 2002).
- The Southern California Association of Governments (SCAG) used PUMS to analyze journey-to-work patterns by socioeconomic characteristics such as race and income level and studied whether socioeconomic backgrounds caused significant differences in travel times (they did not), whether transportation mode made large differences in travel time (almost 75% of transit users incurred more than 30 minutes travel time to work, compared with fewer than 40% of auto users); and whether low-income and minority commuters relied more upon public transit (low-income commuters were four times more likely to take public transportation than high-income commuters) (FHWA 2002).

MTC has also used PUMS data in environmental justice analyses to understand the overlap among elderly, disabled, minority, and poverty populations (C. Purvis, personal communication, Jan. 2011). The Venn diagram shown in Figure 14 illustrates how the PUMS data can be superior to Census and CTPP tabulations. Straightforward American FactFinder tabular data would provide population percentages for the groups in question, but would not provide information about how the groups overlap. CTPP provides many special cross-tabulations, but often will not have the full set of cross-tabulations needed to understand all the interrelationships. The PUMS data enable users to analyze all the interactions among many variables at once. In this example, PUMS can provide estimates of the population by poverty, disability, minority, and elderly statuses, all defined by the user for the specific analysis needs. These estimates can be cross-tabulated so all the combinations of these variables can be accounted for. Note that for the analysis of minority groups, the PUMS data allow users to combine the separate race and Hispanic origin Census questions in ways that best suit the particular environmental justice analyses being conducted.

Gender and Transportation

PUMS data are often used to study a specific subgroup or to compare subgroups. Some of the recent research and analyses on travel behavior differences by gender have been conducted using PUMS.

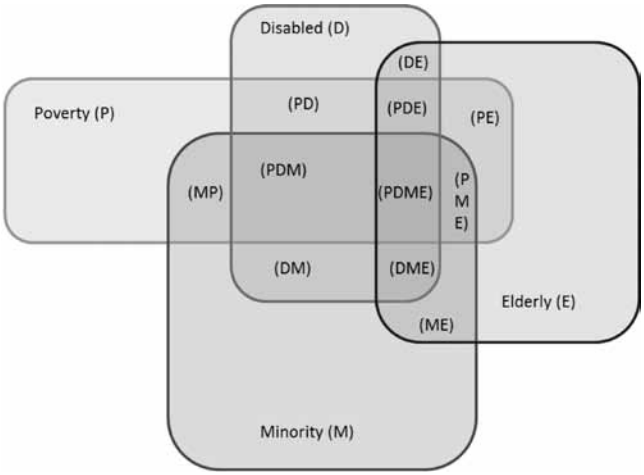


FIGURE 14 An illustration of PUMS cross-tabulation analyses. *Source*: Based on Purvis (2011).

Weinberger (2007) used the 1990 and 2000 PUMS data for the nine-county Philadelphia region to analyze gender differences in commuting patterns. PUMS records were retained for men and women employed in the regular civilian labor force who both reside and work within the study area, and gender differences were assessed along several dimensions:

- Labor force participation and education,
- Commuting travel time,
- Place of residence and place of work geography, and
- Travel time by income and family structure.

In addition to the male-female differential, average journey-to-work travel times were analyzed by race/ethnicity, residential density, household financial responsibility, and a characteristic of industry referred to as sex-based dominance. These characteristics were developed by combining and recoding available PUMS variables. Race/ethnicity was constructed from two Census variables describing race and ethnicity, so commuters were assigned to one of five mutually exclusive categories: non-Hispanic White, Black, Asian, Hispanic, and Other. Residential density was defined dichotomously as urban or suburban, depending on PUMA density. The income burden variable was developed as a proxy to measure household financial responsibility, and was defined as the individual's income divided by the household's income. Finally, the industry variable was reclassified into one of three groupings: male dominated, female dominated, or neutral. The classification rule considered the proportions of men and women in the labor force and compared it to the proportions of men and women in the industry under consideration. If the ratio of the percentage of women in an industry to the percentage of women in the labor force exceeded the threshold of 1.25, it was considered female dominated; if that same ratio was below 0.75, the industry was considered male dominated. Otherwise, it was considered neutral (Weinberger 2007). This study highlights a key advantage of using the PUMS data: the ability to recode and combine variables to be as meaningful as possible for analyses. PUMS data users are not confined to analyses based on prespecified tables and table categories.

In another study, Krizek et al. (2004) used several data sources, including the year 2000 PUMS data, the 2001 NHTS data, and the year 2000 Twin Cities Travel Behavior Inventory survey data, to uncover gender differences in cycling across three dimensions:

- The overall frequency of all cycling trips,
- Commute-only behavior, and
- Cycling behavior of urban versus suburban residents by gender.

Since each of the data sources had limitations in its ability to shed light on the research questions, the authors used all three to provide a comparative picture of relevant differences. Although PUMS and the Travel Behavior Inventory showed similar differences in the prevalence and duration of cycling commutes of men and women, the two sources' estimates were significantly different (Krizek et al. 2004).

Immigration and Transportation

Over the past several years, Blumenberg and others have investigated the travel behavior of immigrants and its current and future effects on transportation policies and programs. This research has relied on a wide range of data sources, including ACS, NHTS, and PUMS data. Because the research examined the specific subpopulation of immigrants, data were not always easy to obtain. The researchers performed cross-tabulations of the PUMS data to better understand the relationships between travel, immigration, and ethnicity. They have also developed discrete choice models to examine commuting characteristics.

Blumenberg and Shiki (2007) used data from the 2000 PUMS to examine the commute mode choice of California's foreign-born population and, more specifically, the relationship between length of U.S. residency and transit usage rates, controlling for other factors likely to influence mode choice. They also used data from the 2000 PUMS to develop discrete choice models to examine the carpooling behavior of foreign-born workers in California relative to solo driving, public transit, and walking (Blumenberg and Shiki 2008).

Blumenberg and Evans (2010) and Blumenberg and Song (2008) drew on data from the 1980, 1990, and 2000 PUMS to further describe immigrants' travel patterns in California, focusing on commute mode choices. They found that immigrants rely more extensively on alternative commute modes (carpooling and transit) than native-born commuters, but over a relatively short period of time in the United States, immigrants assimilate away from these alternative modes and increasingly rely on solo driving. The PUMS analysis

was part of a three-component project that also included focus groups with Mexican immigrants and interviews with community-based organizations.

Blumenberg and Song (2008, p. 58) summarized the strengths and weaknesses of using PUMS:

> The benefits of using PUMS derive from the large sample size for California as well as the inclusion of detailed demographic information such as race/ethnicity, immigrant status, and year of arrival. The data have some limitations, however. The most significant perhaps is the lack of information on travel other than for the commute. The survey asks respondents how they 'usually' traveled to work during the week prior to the survey and, in doing so, precludes data on trips unrelated to commute; non-work trips comprise a significant portion of travel and exhibit a different mode distribution than commute trips.
>
> Another drawback lies in the lack of detailed information on respondents' residential location, given studies show a relationship between the residential location and travel behavior. Finally, PUMS data are cross sectional; although we disaggregate the data over time (across three census years) and by how long immigrants have lived in the U.S., we are unable to follow individuals over time. Still, the PUMS is the best source of data for examining immigrant travel behavior given the lack of other relevant, large sample survey data to examine the travel of immigrants.

Several other researchers have also relied on PUMS data to analyze immigration and transportation. Myers (1996) used data from the 1980 and 1990 PUMS to study immigrant use of public transit over time. Purvis (2003) used the 2000 PUMS for a similar analysis of immigrant transit use in the San Francisco Bay Area. McGuckin and Srinivasan (2003) used 1990 PUMS data to study the relationship between length of time in the United States and auto ownership levels.

Analyses of Jobs Access

PUMS data have been used to support a variety of transportation policy assessments, including jobs access program policies and land use policies.

Thakuriah et al. (2005) created a profile of participants in the FTA's Job Access and Reverse Commute (JARC) program. The authors analyzed participant data for 23 locations across the country. To contextualize their analyses, they used Census 2000 PUMS data to compare and contrast JARC service riders with automobile and transit commuters in terms of socioeconomic and household characteristics. The PUMS data enabled them to make the comparisons against a common set of data and avoid having to assemble up to 28 different commuter profiles.

Hu and Giuliano (2011) compared the job accessibility of low-income and high-income job seekers over time. For their analyses, they relied on census tract-level demographic and employment data from the 1990 and 2000 Decennial Censuses. However, the available Census tabulations at the tract level did not allow the authors to analyze people by age, job type, and income, so they relied on the year 1990 and year 2000 5% PUMS data for 92 PUMAs for these multiway classifications. They applied iterative proportional fitting (IPF) techniques to assign the PUMA-derived intervariable relationships to the tract-level marginal totals of households by income, persons by age, and workers by job type. As the authors note, the assumption that the basic relationships between the variables measured at the PUMA level is maintained for smaller geographic areas (in this case, tracts) affects analyses by reducing the variance of variables. Nevertheless, this assumption is commonly made. The resulting research was able to measure differences in job accessibility over time and spatially.

In another analysis of jobs access, the University of Wisconsin–Milwaukee Center for Economic Development (2004) used 2000 PUMS data to examine vehicle ownership by income levels. The researchers estimated both the percentage of households above and below the poverty level and families above and below the poverty level. Because the PUMS data include nearly the full Census records, the PUMS data allowed the researchers to account for the subtle but important definitional difference between families and households with unrelated members that were not captured by available Census tables.

Land Use Policies

Deal et al. (2009) analyzed whether growth management policies affect commuting mode choice and transit use by comparing 95 metropolitan areas across the United States. The authors used several different data sets, including 2000 Decennial Census data, 2005 ACS data, and year 2000 and 2005 PUMS data sets. PUMS provided sample data for several housing unit variables, such as year of construction and commuting modes of household residents. These variables were used to derive development location indexes for individual regions, and to analyze potential causal mechanisms for transit commuting. Sixteen metropolitan statistical areas (MSAs) and primary metropolitan statistical areas (PMSAs) in states with growth management policies were compared with 79 MSAs and PMSAs in states without growth management policies. Comparison of the 2000 and 2005 PUMS data showed a 0.47% increase in transit commuting in the growth management areas and a 0.10% decrease in transit commuting over that time in the non-growth management areas. In addition, the PUMS data indicated statistically significant differences in occupancy rates between the two groups. The authors also relied on the PUMA geographic definitions to provide convenient subregional zones to develop indices of new development and to determine how the development shifts broadly map against transit availability. Based on their research, the authors conclude that the increased transit use for commuting in growth management areas is more likely

to be the result of lower housing vacancy rates in the growth areas, rather than the result of redevelopment that favors transit-friendly subareas.

Haas et al. (2006) performed an analysis of housing and transportation cost trade-offs for 28 U.S. metropolitan areas. To support the development of a transportation cost model, they needed precise estimates of income levels. They used the year 2000 Census 5% PUMS to estimate the weighted average income of households in specific income categories. For instance, to determine what actual income to use in the income bin range of "Less than $20,000," they used the PUMS data, which provide a count of households at each income level. By querying the PUMS data for households by income restricted to just households earning an income of $0 to $20,000, and to households not living in group quarters, they could identify that the weighted average income in that bin and in one PUMA was actually, $10,385 for all households, $9,837 for renters, and $11,368 for owner households (Haas et al. 2006). They performed the analysis for each PUMA in 28 metropolitan areas. Although they acknowledge that the approach was compromised by the large sizes of PUMAs (they were performing analyses at the tract level, and needed to assign the PUMA-weighted averages to all constituent tracts), they believed that the PUMS-based approach was an improvement over a naïve midpoint approach (Haas et al. 2006).

Finally, in an evaluation of transit capitalization opportunities and benefits in the San Diego region, Duncan (2008) compared demographic and commuting attributes of households that own a condominium with those of households in a single-family home. Based on the measured differences, Duncan inferred that the condominium market segment has some additional share of potential buyers interested in station-area housing. The detailed housing type information in PUMS allowed for this comparison and conclusion. Even so, the study noted that PUMS-based analyses were limited by the geographic limitations of PUMS, and the analyses could have been improved with additional data items not contained in PUMS (e.g., nonwork travel information, attitudinal/lifestyle preference information) (Duncan 2008).

PUBLIC USE MICRODATA SAMPLE TO SUPPORT TRAVEL SURVEYS

MPO and state DOT planners use PUMS data to support travel surveys in a variety of ways, including in the development of survey sampling plans, weighting of survey results, and the validation of survey results.

Public Use Microdata for Sample Planning

For most agencies, the most significant travel survey effort is the periodic household travel and activity survey, in which all members of a sample of households in a region are asked to complete one-day or multiday travel diaries. The results of these surveys are usually used as key inputs into travel demand modeling efforts.

Household travel surveys usually rely on stratified sampling strategies that aim to ensure that an adequate sample of households with prespecified characteristics are included in the survey. Often, controls are set by some combination of characteristics, such as household size, household auto availability, household workers, and/or household income categories. The stratification targets are usually developed with CTPP tabulations or directly from Census data tabulations, but PUMS data are sometimes used in the sample planning and (more frequently) the survey weighting.

For the recent large-scale Michigan Statewide Household Travel Survey, PUMS data were used to ensure adequate representation of all regions and subregions within the state (Faussett 2006). Census 2000 PUMAs were used as the basis to allocate households. For each PUMA, the households were summarized by size, auto ownership, and number of workers. Then, the percentage of households by region was determined for each PUMA. With the percentages of households by region for each PUMA determined, the number of households per cell for each region within each PUMA was calculated.

PUMS data have also been used to design sample stratification (oversampling) strategies (web-based survey scan 2011). The most significant challenges of household travel surveys are reaching "difficult-to-reach" communities and gaining their cooperation to participate in the surveys. Household travel surveys can often have problems obtaining samples that are representative of the overall household population in terms of age and race/ethnicity. In addition, to ensure robust travel demand models, survey planners often seek to oversample households that have no vehicles available, but in many regions these households are rare and hard to reach. Analysis of PUMS data to identify the range of characteristics of these households can suggest strategies for targeting them. The PUMS data are especially useful for this purpose because data users may separate household records from group quarters records and because the PUMS data provide a full range of Census characteristics. However, the lack of geographic specificity limits the ability of survey planners to use these analyses to identify specific locations to target.

Public Use Microdata Data for Survey Weighting

Once the survey data are collected, post-stratification weights are usually applied using the sample control variables to address any deliberate oversampling or undersampling. Once weights are applied, the weighted survey results will match the population targets for the control variables.

To weight the survey results to account for deliberate and circumstantial oversampling and undersampling, analysts factor all the survey results with a specific characteristic by the ratio of the sum of the actual households with the characteristic (taken from Census-based estimates) to the sum of survey records with the characteristic. When multiple characteristics are used to expand the data, the weighting for each characteristic changes the weighting for previous characteristics. Consequently, analysts often use iterative proportional fitting, or "raking" techniques, to establish weights that try to account for each characteristic. IPF generally allows users to establish record weights that enable several control variable categories to be sized to simultaneously match the available Census estimates for the region.

PUMS data are well suited to provide the Census estimates in this process because they can be summarized for any combination of control variables. CTPP tabulations and direct ACS/Census tabulations can also be (and frequently are) used for the weighting control totals, because surveyors often use control variables that are well represented in the predefined tabulations (e.g., household size, workers, vehicle availability, income).

The recent Ohio DOT GPS Household Survey is an example of a survey for which PUMS data were used in weighting (Ohio DOT Consulting Team 2011). The survey team employed a multidimensional stratified random sampling approach. Households were stratified on the basis of geographic characteristics and household characteristics:

- Geographic characteristics
  - To ensure adequate representation from all areas within the study area, some counties were oversampled.
  - To support future travel demand modeling uses of the data, certain block groups were oversampled, including—
    - Areas near major universities, and
    - Areas deemed to have higher propensity for transit usage.
- Household characteristics
  - To support modeling that relates trip-making to household characteristics, targets were set for cross-classified household types, defined by—
    - Household size (one person, two people, three people, and four or more people);
    - Household workers (zero, one, two, and three or more workers);
    - Household auto availability (zero, one, two, and three or more autos available);
    - Annual household income (less than $25,000; $25,000 to $50,000; $50,000 to $75,000; more than $75,000); and
    - Household life stage (adult only households, households with children present, retiree households, adult student households).

To account for this targeted oversampling, the Ohio DOT (ODOT) consulting team employed a weighting strategy that used IPF techniques to match weighted household survey summaries to similar summaries of the region's best available data sources for the geographic and household characteristics used in the sampling. The choices of the best data sources to use in the weighting were based on the currency of the data (data pertaining as closely as possible to the survey period were sought) and availability of household characteristics weighting variables (Ohio DOT Consulting Team 2011).

A combination of data sources were used for the weighting:

- Geographic targets from 2010 Decennial Census PL94-171 data tabulations, and
- Household characteristics targets from the 2005–2009 ACS Public Use Microdata Sample (5-year ACS PUMS data).

The new 2010 population data were the most recent population and household estimates available, so they were used to address geographic undersampling/oversampling.

However, because the 2010 data did not include household characteristics data, and because standard ACS tabulations did not allow the survey team to summarize the households by all of the household characteristic variables, the team needed to rely on PUMS data instead. Using the PUMS microdata with sampling weights enables custom tabulation of household characteristics, so the data can be summarized by variables such as the ODOT household life stage (Cambridge Systematics 2011).

The 2005–2009 ACS PUMS data were processed as follows (Cambridge Systematics 2011):

- The PUMS data for PUMAs within the survey study area were obtained from the Census Bureau website. Because the study area did not match PUMA definitions exactly, PUMS household weights were adjusted for records from PUMAs that included areas outside the study area based on the population proportion of the PUMA in the study area. This enabled the weighted PUMS records to properly reflect the study area population.
- The survey team retained housing and person records only for nonvacant housing unit records (no group quarters records). The data set for weighting included 41,428 individual household records, representing 765,448 PUMS-weighted households.
- The survey team recoded PUMS person variables and summarized them to their households to match household survey characteristics categories.
  - PUMS person age variable categorized into three dummy variables (less than 18; 18 to 64; 65 or more),
  - PUMS school attendance variables categorized into two dummy variables (K-12 student; college student),

  - PUMS employment status variable categorized into one dummy variable (Employed),
  - Adult student and adult full-time student dummy variables created based on age, student, and employed dummy variables, and
  - All person-based dummy variables summed by household, and sums merged to household records.
- The survey team recoded PUMS household variables to match household survey characteristics categories.
  - Per Census Bureau recommended practice for ACS PUMS, household income variable adjusted to account for multiple data collection years using PUMS adjustment factor,
  - Adjusted household income categorized into four categories (less than $25,000; $25,000 to $50,000; $50,000 to $75,000; more than $75,000),
  - Household size (sum of people in household), workers (sum of employed persons in household), and vehicles categorized into four categories each, and
  - Households assigned to one of four life cycle categories (adult households; adult student households; retiree households; households with children).
- The survey team combined the PUMS household variables and the summarized person variables to summarize the PUMS records by household type definition.
  - Households with children—Any two or more person household with at least one household member less than 18 years old,
  - Retiree households—Any household with at least one household member 65 years old or more and with no employed people,
  - Adult student households—Households not qualifying in the previous categories and with—
- All household members are adult students; or
- Two household members with at least one a full-time adult student; or
- Three household members with at least two students and at least one full-time adult student; or
- Four or more household members with at least half being adult students.
  - Adult households—Households not assigned to one of the existing categories.

Once the original PUMS data file had been modified to include the additional variables that correspond to the household survey definitions, the survey team applied IPF techniques to establish the survey weights. They compared the household survey control variable category sums with summaries of the PUMS file, and then collapsed control variable categories to enable effective IPF application. They developed marginal totals from the PUMS file and tabulated household survey data to form the joint distribution targets. Finally, they performed the IPF procedures. The resulting set of survey record weights resulted in data summaries that closely matched the PUMS household characteristic summaries (Cambridge Systematics 2011).

As an additional step in the weighting, some analysts have used the PUMS data to evaluate how well the survey matches the population on uncontrolled variables (Nilufar 2003). This step can be used to define additional weighting controls or simply to understand any important differences between the collected sample and the overall study area household population. For instance, after weighting on household variables, analysts may want to compare the age distribution or ethnic mix of weighted household survey respondents to that of the overall household population for the study area.

Using IPF techniques to weight on a combination of household variables and person variables can be problematic, but Konduri et al. (2009) developed a method for estimating household travel survey weights that can consider both household characteristics and the characteristics of people within the households to develop weights that are consistent. This entropy optimization method was developed using year 2000 Census PUMS data for Maricopa County, Arizona, and has been extended for use in synthetic population analyses, as discussed here.

PUMS data have also been used for weighting other survey data beyond household travel survey weighting efforts. Gao et al. (2008) investigated public opinions toward car sharing and the latent demand for these services through a survey of Austin, Texas, residents. The survey included both hard copy and web-based versions, and it was preceded by an advertising campaign in and around the University of Texas campus. Because of the survey procedures, the completed surveys did not reflect the overall population of the Travis County study area. For instance, students were significantly overrepresented, as were people with higher educational attainment levels. Consequently, the researchers used the 2005 Census PUMS data to weight the survey results along four dimensions. The PUMS data and the survey data were summarized in four-way cross-tabulations of—

- Student status (yes or no),
- Education level (associate's degree or less, versus bachelor's degree or higher),
- Age (18–35 years old, 36–55 years old, and 56 or older), and
- Annual household income (less than $25,000 per year, $25,000 to $75,000 per year, and more than $75,000 per year).

The respondent-specific weights were calculated to be the normalized ratio of PUMS probabilities to sample probabilities. The weighted survey results were then used in all the statistical analyses of the survey data. The researchers concluded that the weighted survey results "provided rich information on public opinion of different aspects of the [car sharing] program, as well as the expected demand on the service and possible changes in travel patterns" (Gao et al. 2008, p. 1).

One issue that has arisen with using PUMS data to expand travel surveys is that estimates derived from the PUMS data may be somewhat different than estimates based on other Census data products. Therefore, the application of weights to have travel surveys better match PUMS data summaries may lead to survey results that are inconsistent with other summaries of Census data. In her master's degree work reported in the CTPP Status Report, April 2009, Laura McWethy compared using PUMS and an IPF technique versus using CTPP (http://www.fhwa.dot.gov/planning/census_issues/ctpp/status_report/sr0409.cfm).

In a preliminary review of travel survey results, one agency compared published Census estimates with weighted survey results that were weighted with PUMS data, and found the following estimates for average household size:

- 2005–2007 3-year "Selected Social Characteristics" profile reports an average household size of 2.32 ± 0.03;
- 2007–2009 3-year profile reports an average household size of 2.36 ± 0.03;
- 2009 1-year profile for the PUMA area reports an average household size of 2.39 ± 0.05;
- 2009 1-year profile for the equivalent urbanized area (defined to be the same as the PUMA) reports an average household size of 2.38 ± 0.06; and
- The weighted travel survey summary reports the calculation of an average household size of 2.22 (S. Payne, personal communication, Aug. 2011).

Thus, there is a meaningful difference between the PUMS-weighted household survey results and the published Census estimates. The survey team is currently assessing the reasons and ramifications of the difference on future analyses, so it is not certain whether the problem lies in the PUMS data use (S. Payne, personal communication, Aug. 2011). However, whether or not the use of PUMS contributed to the specific issue described, PUMS data users need to be aware of these possible differences. In each year, the ACS data are collected for about 2.5% of the population, and the PUMS data represent only about 40% of the ACS data records. Therefore, both estimates derived from the ACS and estimates derived from the PUMS data are subject to significant sampling error. The PUMS' sampling error will be larger, so estimates based on PUMS are less reliable. In addition, the PUMS data are subject to data swapping and other confidentiality protections, which could also lead to differences between ACS and PUMS estimates.

Even when PUMS data are not used in analyses, PUMA geographic definitions can still be helpful. The Chicago Metropolitan Agency for Planning (CMAP) recently completed a comprehensive travel and activity survey for northeastern Illinois (data collection performed in 2007 and 2008). This travel diary–based survey of more than 10,000 households was performed to support regional travel demand modeling, and relied on a stratified sampling plan that was similar, in general terms, to most other regional household travel and activity surveys.

The survey team performed the survey weighting by using 2005–2007 ACS tabulations. Six ACS variables were used to support an IPF-based weighting approach. The survey team sought to apply the weights for subregional districts because of known differences between travel patterns in the region. The CMAP survey team identified the PUMS geography for PUMAs as a practical and useful delineation of districts. Figure 15 shows the PUMS-based subregions for which weighting procedures were completed. The direct relationship between PUMAs and Census tracts enabled the survey team to summarize the ACS data by PUMA easily. As noted in the survey weighting documentation, "using eleven zones was a compromise between having many smaller zones which would have more similarities in travel patterns and keeping the sample sizes large enough so that the survey data could be balanced and weights remain within a reasonable range" (CMAP 2009, p. 8).

FIGURE 15 CMAP's PUMA-based weighting sub-regions.
Source: CMAP (2009).

Public Use Microdata Data for Survey Validation

PUMS data can be used to review the representativeness and validity of survey data after they have been collected or even while they are being collected.

MWCOG used PUMS commuter data to perform validation tests on its household survey data (R. Griffiths, personal communication, Apr. 2011). First, it summarized the PUMS data for the PUMA-to-POW-PUMA commuter flows and means of transportation for workers in households. Using the PUMS data enabled MWCOG to separate household workers from group quarters residents, which was desirable because group quarters residents were not included in the survey sampling frame. When survey data became available, MWCOG staff was able to compare them with the PUMS summaries along these dimensions. Because the PUMA and POW-PUMA geographic areas are so large, the geographic comparisons were at broad areas only, but this level of information is still useful for multistate/multijurisdictional areas such as Washington, D.C. Significant differences between the PUMS estimates and the survey estimates were used to identify potential survey sampling issues, such as within-county geographic biases in the sampled households and biases in the types of households participating in the survey (R. Griffiths, personal communication, Apr. 2011).

Preliminary geocoding of survey records will generally suffice for assigning work trips at the PUMA-to-POW-PUMA level of geography, so the PUMS-based analyses can provide a quick check on the survey home-to-work trip geography without waiting for manual geocoding efforts to be completed. This means that household survey teams could incorporate this type of PUMS-based data review into their set of tests that they perform to monitor survey efforts as they are being undertaken.

Pearson et al. (2009) used tabulations of the Census 2000 PUMS data to evaluate the potential effects on trip estimates of excluding households without telephones from a household travel survey in the Lower Rio Grande Valley. The authors tabulated the PUMS data for the region by household income, household size, and telephone availability. They developed an estimate of the number of nontelephone households from the PUMS data, and then hypothesized different levels of potential trip-making by nontelephone households. The trips generated under the "worst case" scenario for nontelephone household trip-making were then compared with the confidence limits of the trip estimates derived from the regional survey. For their analysis, the potential trip-making of nontelephone households, which comprised about 6% of Valley households, fell within the 95% confidence interval of the telephone survey. So it was concluded that the lack of nontelephone surveys did not materially affect the survey accuracy, and remedial data collection efforts were unnecessary.

PUBLIC USE MICRODATA SAMPLE DATA TO SUPPORT TRAVEL DEMAND MODELING EFFORTS

State and local transportation planners rely on travel demand models to support several mandates, including—

- Development of long-range transportation plans to guide capital and operating investment decisions;
- Analysis of potential air quality conformity ramifications of alternative scenarios; and
- Evaluation and environmental review of potential projects.

The travel demand models developed for metropolitan areas and states differ in complexity and capability. *TRB Special Report 288: Metropolitan Travel Forecasting: Current Practice and Future Direction* describes the "state-of-practice" of travel demand forecasting in the United States (TRB 2007). The special report summarizes the common "four-step trip-based" framework on which most model systems are based, and discusses emerging requirements and uses of the model systems.

The traditional trip-based modeling framework includes the following elements (Donnelly et al. 2010):

- **Model input preparation** – Development of highway networks, transit networks, and estimates of households and employment by zone;
- **Household submodels** – Summary of households by their characteristics (e.g., household size, income);
- **Long-term travel behavior submodels** – Estimation/prediction of household auto ownership;
- **Trip generation model component** – Estimation/prediction of the number of trips by trip purpose;
- **Trip distribution model component** – Estimation/prediction of the trip origins and destinations;
- **Mode choice model component** – Estimation/prediction of travelers' mode choices for their trips;
- **Time-of-day/peak spreading submodels** – Estimation/prediction of the temporal distribution of the trips; and
- **Assignment model component** – Estimation/prediction of auto and transit volumes and travel times on networks.

Travel demand models are developed from several different input data sets, including different kinds of surveys, count data, socioeconomic and land use data, and geographic data. Decennial Census, ACS, and CTPP data are often used for validating models as well.

PUMS data are less useful for many of the model components that relate to specific trips because of the lack of geographic detail afforded by PUMA definitions and the lack of specificity in the PUMS commuting geography. For instance, the Texas Transportation Institute works closely with more than 20 small MPOs to develop the demographic data needed for travel demand modeling and other transportation planning studies, but usually does not use PUMS data because the 100,000 minimum population requirement for PUMAs results in only one or two PUMAs in the smaller MPOs, and the geography usually does not correspond to the

MPO planning area boundaries. It is much simpler, in their view, to work with available data tabulations for smaller Census geographic delineations that better correspond to the MPO geographies (MPO area and traffic analysis zone structure) (P. Ellis, personal communication, Aug. 2011).

Public Use Microdata Data Use in Travel Model Calibration

Despite the geographic limitations of PUMS, modelers have taken advantage of PUMS data to develop household submodels and long-term travel behavior model components. Some of these model components and submodels can be developed at regional and subregional levels, so the PUMS microdata provide an excellent means for development of modeling relationships.

Because PUMS allows special tabulations that are not normally available for a metropolitan area, county, or state, many planning agencies use PUMS to understand household structure and to prepare inputs for the first major step of the traditional travel demand forecasting process, the trip generation model component. In a spring 2005 review of statewide travel demand models, 7 of 32 statewide model calibration efforts relied in part on the use of PUMS data (Horowitz 2006).

Household Characteristics Model Development

MTC was one of the early users of PUMS data for transportation demand models. It used PUMS data from as early as 1980 to develop household submodels. The PUMS data were used to improve the previous household submodels, which had failed to adequately capture interrelationships of household income levels, workers, and size. Analyses of the PUMS data demonstrated that higher-income households with workers had higher incomes than higher-income households without workers. Similarly, higher-income households with workers had larger household sizes (and higher incomes) than higher-income households without workers. MTC applied PUMS-based county adjustment factors to address these issues (C. Purvis, personal communication, Jan. 2011). MTC also used PUMS 2000 data to develop estimates of past and future student status (nonstudent, elementary, high school, college) by age. These estimates were used to calibrate the school enrollment submodel (C. Purvis, personal communication, Jan. 2011).

Portland Metro uses PUMS data as a basic input into its household composition forecasting process. Independent forecasts of household variables, such as household sizes, age distributions, and household income levels, are made using economic forecasting processes, and the PUMS data are used to provide the distributions of households subject to these marginal estimates. The PUMS data provide the ability to define discrete categories, such as 16 income categories, and to consider variables that may not be available from standard tabulation, such as consideration of both personal income and household income (D. Yee, personal communication, Apr. 2011).

Englund et al. (2010) used the 2005–2007 and 2006–2008 3-year ACS PUMS data to develop a means to better reflect worker characteristics in CMAP's existing travel demand model. The authors used the PUMS data to review the characteristics of workers in the region, and to demonstrate how the model system could be enhanced. ACS PUMS cross-tabulations of commuting mode, reported commuting travel time, worker earnings (rather than household income, which had been used in previous modeling related analyses), and occupation (Standard Occupation Classifications established by the Bureau of Labor Statistics) were developed and analyzed (Englund et al. 2010). Based on this review, the authors developed new worker submodels and corresponding revised trip generation models that allow for two worker earnings levels. The ACS PUMS data were used to develop a cross-classification model that estimates the probability that a worker is a high-earnings worker, given household characteristics, number of workers, adults, children, and household income quartile. The probabilities were then built into the CMAP trip production and trip attraction components within the trip generation model (Englund et al. 2010).

Englund et al. (2010) also used CTPP data to propose improvements to the model's trip distribution and mode choice components. Although they did not rely on PUMS data directly for these analyses, they used the PUMA definitions to define districts.

Vehicle Availability Model Development

Several years ago, Purvis (MTC) explored the usefulness of the 1990 PUMS data set as a basis for estimating logit choice models of automobile ownership for the Bay Area (Purvis 1994). He demonstrated the consistency of logit choice models based on PUMS and household survey data, concluding that the PUMS data are useful for metropolitan areas and states that do not have access to recent household travel survey data. Following this initial work, similar auto ownership models that relied on PUMS data were developed in other regions throughout the country, including New Hampshire, Philadelphia, Honolulu, Atlanta, Kansas City, and New York City (Cambridge Systematics 1996, 1997; Ryan and Han 1999). These studies also demonstrated the effectiveness of the PUMS data for auto availability modeling.

Purvis and the other authors noted the major weakness of this approach: an inability to include zonal variables or accessibility measures in the models because the individual households provided in the PUMS data are not identified by their location except at the level of districts including at least 100,000 persons. Thus, the PUMS data source was characterized as a "second best" data set for automobile ownership model development, one that cannot completely substitute for

data from a comprehensive household travel survey. Baber (2004) found that neither PUMS data nor NHTS data on their own provided enough specific land use and built-environment information to accurately capture auto availability. However, even when household travel survey data are available and used for auto ownership model estimation, the PUMS data are frequently used to validate the estimated models.

Internal-External Model Development

MWCOG used PUMS data to help develop the external model component of its travel demand model. MWCOG staff assembled and analyzed national PUMS data to identify workers with work locations in the MWCOG model region. The PUMS data enabled MWCOG to estimate the number of out-of-town workers who commute to the model region and the number of out-of-town workers who reside in the model region on a nonpermanent basis. This analysis helped MWCOG staff better understand seeming inconsistencies in the region's jobs-housing balance (R. Griffiths, personal communication, Apr. 2011).

Assessment of PUMS for Model Calibration

The primary reason that PUMS data are not used extensively in travel demand model calibration is that the PUMA geography does not support the level of spatial detail that most models require. Often, modelers need to make explicit or implicit assumptions about how the intervariable relationships measured at the PUMA level apply for individual traffic analysis zones (TAZs) or tracts that comprise the PUMA. Sometimes, PUMA average values or modeled relationships developed at the PUMA level are assigned to all the smaller geographic delineations within the PUMAs. Another often-used approach is to apply IPF techniques that rely on the PUMS data to form a joint distribution matrix and that rely on other Census data products (such as ACS) for small area marginal totals. With these inputs, small area joint distributions are developed that adhere to the marginal totals and reflect the underlying PUMS joint distribution. The application of IPF is described further in the population synthesis discussion.

Travel Demand Model Validation

One other common travel demand modeling application of the PUMS data is for model validation. The PUMS data provide the means to tabulate subregional estimates of household characteristics, including information on commuting. The FHWA Model Validation and Reasonableness Checking Manual and other modeling guidance suggest that modelers should use the PUMS data to ensure that base year cross-classifications are consistent with expected values (Barton Aschman Associates and Cambridge Systematics 1997). Some modelers also compare commute trip geography indicated by the travel demand models to the PUMS' PUMA-to-POW-PUMA flows (web-based survey scan 2011).

MTC used the 2000 PUMS data to develop summaries of county-to-county commuters by means of transportation by household income quartile, and used these outputs to validate the travel demand model's work trip distribution. Often, these types of validation exercises can be accomplished with CTPP rather than PUMS, but sometimes PUMS data will be substantially more current (CTPP being available only periodically) and can provide more flexibility in variable definitions. For MTC, the 2000 CTPP did not provide the income levels that were needed for direct comparison to the model, so the PUMS data (which can be summarized by any income delineations) were used instead (C. Purvis, personal communication, Jan. 2011).

Special Demand Modeling Applications of Public Use Microdata Data

Travel demand models and modeling techniques are being adapted and improved to address several new analytical needs, including the ability to address the following:

- Motor vehicles emissions and speeds,
- Induced travel,
- Land use policies,
- Nonmotorized travel,
- Transportation policies,
- Cumulative and secondary impacts,
- Environmental justice,
- Economic development,
- Planning for emergencies, and
- Changes in population demographics.

The TRB *Special Report* notes that these emerging factors have resulted in a need for models that are—

(*a*) more completely specified, to address more variables of interest;
(*b*) more disaggregate in time, space, and categories of activities; and
(*c*) better able to account for supply-side effects (traffic operations) (2007).

Many modelers have enhanced their conventional four-step trip-based models in response to these additional requirements, and some have chosen to invest in more advanced modeling practices. In addition, analysts are using modeling techniques to perform special analyses outside the confines of the regional model framework. *Special Report 288* outlines some of the new conceptual approaches and analytical techniques that are being taken in modeling, and *NCHRP Synthesis 406: Advanced Practices in Travel Forecasting* provides more detail on travel demand forecasting innovations that are being implemented in a number of metropolitan areas (Donnelly et al. 2010).

The Advanced Practices in Travel Forecasting Synthesis and several other sources describe the recent development

history of advanced travel demand models, beginning with trip-based models and proceeding to tour-based models, TRANSIMS (TRansportation ANalysis SIMulation System), and activity-based travel demand models (Vovsha et al. 2005; Bowman 2009a; Donnelly et al. 2010).

Among the important model developments has been an increased consideration of land use policies in a transportation modeling framework. Many growing regions must consider options other than transportation capital improvements for addressing future mobility needs. Their MPOs, therefore, need to be able to model land use policies such as increases in overall density, urban growth boundaries, intensification around rail stations, and more mixed housing and employment. Models must be sensitive to these variables (TRB 2007). Researchers have found the PUMS data to be particularly valuable in transportation–land use interaction modeling.

The Transportation Economic and Land Use Model (TELUM) is the successor to one of the pioneering integrated land use models, DRAM/EMPAL. It is distributed under FHWA sponsorship, along with TELUS (Transportation, Economic & Land-Use System), a decision-support system that helps MPOs and state DOTs manage annual transportation improvement programs and carry out other agency responsibilities. As an aggregate land use model, TELUM does not rely on population synthesis as do the land use models discussed here. Nevertheless, TELUM users (primarily small and medium-sized MPOs) are directed to use PUMS data to obtain an important set of input data (TELUS 2005).

A key concept within TELUM is that as the regional mix of employment types varies, so does the region's household income distribution. A TELUM employment module (TELUM-Emp) forecasts employment at places of work by type (e.g., retail, manufacturing). These forecasts are then converted to households by income group at place of work by multiplying the employment forecasts by a matrix of conversion factors. Although the model system provides a default matrix, users are advised to develop region-specific conversion factors using PUMS data. Specifically, TELUM users access the PUMS data for their regions in order to calculate the distribution of the number of heads of household, by income group, employed in each industry. Because household income is a continuous variable in PUMS and there are more than 500 detailed Census occupation codes, TELUM analysts have great flexibility in tailoring the cross-tabulations of income and occupation (TELUS 2005).

Weinberger and Goetzke (2009) used the Census 2000 5 percent PUMS data to develop a joint automobile ownership/residential location model that captures the impact of a person's previous observations and experiences on that decision. The PUMS data were used to identify the characteristics of recent movers currently residing in the metropolitan areas of Boston, Chicago, Philadelphia, San Francisco, or Washington, D.C. The PUMS migration variables allow data users to identify households and people that have moved to their residence in the past year. The multinomial logit models included a wide range of household and head-of-household characteristics, such as household size and household income, as well as the age, gender, race, and educational attainment of the head of household.

The models also took into account the previous residence location of the PUMS respondents. The researchers demonstrated that the previous residence location has a statistically significant impact on auto ownership decisions. People who had moved from cities with lower auto ownership levels were more likely than others to not own a vehicle. Weinberger and Goetzke conclude that people's preferences for low or high levels of auto ownership are learned from previous experiences, and they discuss the policy implications of this self-reinforcing cycle (Weinberger and Goetzke 2009).

Morris and Smart (2011) used Census PUMS data from 1980, 1990, and 2000 for the Los Angeles region to develop models to test whether ground-level ozone pollution levels affect the residential location decisions of either physicians or laypeople. The idea of this modeling effort was to use hedonic price models based on willingness to pay for housing and commuting to examine the differential valuation of reduced pollution levels between doctors and others.

The 1980, 1990, and 2000 PUMS data were screened to include full-time employed homeowners only (194,023 person records) and were reweighted to account for differential sampling rates across the three data sets. Then, in order to assign ground-level ozone levels to each PUMA, the researchers used California Air Resources Board measurement data to develop an ozone level gradient model for the region. The prediction gradient was collapsed onto the PUMA geography to estimate an average predicted ground-level ozone level for each PUMA (Morris and Smart 2011).

Analyses were performed to test the hypothesis that physicians (who are assumed to have a better understanding of the health effects of smog) are more likely to avoid polluted areas. First, simple descriptive analyses were applied. The average ozone levels of PUMAs were compared with the percentages of doctors living in those PUMAs. These comparisons were made for each of the three analysis years. No clear relationship was apparent, so the researchers employed ordinary least squares regression analysis to compare residence locations of physicians and laypeople while controlling for housing-related and demographic factors. The PUMS data included several variables that were used to control for housing differences—household tenure, years present in the housing, age of the housing unit, and number of rooms in the housing unit. In addition, several PUMS demographic variables were used in the modeling,

including household size, age, gender, household income, level of educational attainment, and race/ethnicity (Morris and Smart 2011).

The researchers also developed regression models that considered demographic, housing, and transportation factors in the PUMS data set. With these models, they tested whether physicians would be more willing to accept longer commuting times than laypeople in order to avoid higher ground-ozone levels. The PUMS data were able to provide general (POW-PUMA level) information on workplace locations, along with commuting mode and time-of-day information. The researchers concluded that there was no clear evidence that doctors' willingness to pay for clean air in commute duration differs from laypeople's willingness to pay (Morris and Smart 2011).

Transportation modelers have begun using advanced market research techniques and econometric modeling concepts, including market segmentation analyses, factor analyses, and cluster analyses, to analyze and better understand travel demand. A few of these efforts have relied in part on PUMS data.

Beckman et al. (2008) used the Census 2000 PUMS data for 10 counties in California to investigate spatial, social, and economic determinants of the joint distribution among travel time, mode choice, and departure time for work. They used latent class cluster analysis techniques to identify the primary determinants (within the PUMS variables) of the workplace commute decision-making process.

The researchers stated that the idea behind the latent class cluster analysis was to analyze the patterns in variance across many dependent variables and to identify groups of people with relatively homogeneous behavior. Classification of each person (PUMA record) in a class was then based on the likelihood of class membership. This was done by assuming that a latent (unobserved) variable can be discerned from the data at hand, and this latent variable was used to explain the data variance. The researchers specified a series of models with different categories in their latent variables, and selected the model that best balanced parsimony and goodness of fit (Beckman et al. 2008).

Through the latent class cluster analysis, seven clusters were identified as optimal in segmenting the population pertaining to mode choice, travel time, and leave time simultaneously. The researchers concluded that through use of latent class clusters, market segments were more easily identifiable than previous attempts in the immigrant travel behavior analysis literature. They also noted some limitations of this analysis that were inherent in the Census 2000 PUMS data, including the limited travel variables collected and the geographic specificity of the PUMAs (minimum 100,000 population threshold) (Beckman et al. 2008).

Zhou et al (2004) used a structural equations modeling and cluster analysis approach to help the San Mateo County Transit District (SamTrans) to better understand customer attitudes and perceptions, then to create market segments that reflect and account for traveler attitudes and to identify market segments in the population that can be targeted for new SamTrans services. A customized survey was used to perform market segmentation analyses of SamTrans users and nonusers. Then, PUMS data were used to relate the attitudinal market segmentation results to the population of the SamTrans service area.

According to the authors, the PUMS data have an advantage over Census tabulations because the PUMS files contain individual- and household-level information. Since the market segmentation model was estimated using individual-level survey data, the model could be directly applied to an individual-based data set such as PUMS. The PUMS records with attached market segmentation assignments were then assigned to detailed geographic areas using Census small area tabular data. The study enabled SamTrans to identify the spatial and modal distribution of its service market based on customer needs and to compete more effectively in the target geographic markets addressing the needs of individual market segments (Zhou et al. 2004).

Finally, several researchers have developed demand models to expand upon the immigration transportation research described earlier.

Chatman and Klein (2011) combined several successive 1-year ACS PUMS data sets (2006–2008) to analyze the determinants of transit commutation among immigrants and U.S.-born residents of New Jersey. They developed multinomial logit regression models using the PUMS records and PUMA-level transportation accessibility and density measures. They built on the significant literature on immigrant commuting, some of which is described earlier, by introducing spatial characteristics into their analyses of these commuting patterns. As the authors note, in most research that has taken advantage of the disaggregate nature of the PUMS data, spatial characteristics have played a limited role. The geographic limitations of the PUMS data preclude the use of neighborhood-level variables, so researchers have tended not to include measures of transportation accessibility or density in their analyses. In addition, Chatman and Klein found that the research generally had not considered workplace spatial measures to the extent possible (Chatman and Klein 2011).

Chatman and Klein calculated several spatial variables at the PUMA and POW-PUMA levels, and found that even with these large subregional geographic delineations, spatial measures could be used effectively to model commuting decisions. They calculated the number of bus stops, number of rail stations, population density and employment density in the home subregion, and population and employment den-

sity in the workplace subregion. The multiyear PUMS data set with the merged spatial information included more than 150,000 workers, about 35,000 of whom were foreign-born (Chatman and Klein 2011).

The PUMS data indicated a significantly higher propensity for the foreign-born workers to commute by bus or rail, so the modeling effort sought to investigate the extent to which these mode share differences were the product of spatial factors or demographic factors, including immigrant status. The models included variables on the number of years in the United States, place of birth, citizenship, occupation, home PUMA spatial characteristics (population density, bus and rail transit availability, bus and rail stops per 1,000 persons), and workplace POW-PUMA spatial characteristics (employment density, bus and rail transit availability, bus and rail stops per 1,000 workers). The final model specifications also included demographic control variables available from PUMS, such as household income, age (and age squared), sex, racial category and Hispanic status, education, family size, and presence of children in the household (Chatman and Klein 2011).

Cline et al. (2009) used year 2006 PUMS data for the state of Texas to analyze differences in carpool formation among Hispanic, immigrant Hispanic, and non-Hispanic white commuters. They developed logistic regression models to isolate the influencing effects of socioeconomic, occupational, and geographic characteristics on the propensity to carpool on the journey to work, and to test if differences between the groups in carpooling remain. PUMS provided the researchers with disaggregate data on commute mode choice and potential explanatory variables, including a wide range of household and person characteristics and geographic characteristics that appear in many transportation models, but also with data on respondents' self-identification (as Hispanic or non-Hispanic), place of birth, and age of immigration. The authors note a few limitations of the PUMS data, specifically that the PUMS data set cannot provide fully detailed data on people's immigration status nor on whether people have valid driver's licenses.

Kim (2009) analyzed commuter mode choice differences among nonimmigrants, new immigrants, and other immigrants through the development of multinomial logit models using a sample of data records from the 2006 ACS PUMS national sample obtained from IPUMS. In addition to the immigration status of the PUMS workers in the sample, the models included a wide range of other PUMS data fields as explanatory variables, including age, income, race, ethnicity, gender, educational attainment, disability status, work industry, region, English-speaking ability, age of home, vehicle availability, work hours, and arrival time at work. The availability of the nearly complete set of collected ACS data in PUMS (and IPUMS) enables data users to test a huge number of variable combinations in disaggregate model specifications.

PUBLIC USE MICRODATA SAMPLE DATA TO SUPPORT POPULATION MICROSIMULATION

NCHRP Synthesis 406 catalogs several advanced modeling strategies that agencies are beginning to implement (Donnelly et al. 2010). Among the most important and interesting modeling innovations are two that rely on Census PUMS data to a large extent:

- The development of activity-based microsimulation modeling systems, and
- The incorporation of land use models into the transportation modeling framework.

A third factor in the increased usage of PUMS data by transportation planners is the increased interest and research into survey data transferability. As household travel survey data collection has become more difficult and expensive, transportation researchers have become more interested in developing simulated or model-based survey data.

As they are being forwarded by modelers, these three advanced modeling developments are increasingly relying on the application of population microsimulation techniques to PUMS data. This section briefly describes the modeling advancements, and then discusses population microsimulation using PUMS data (which is common to the three model advancements).

Activity-Based Models

One of the ways 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. This approach represents an advance over traditional travel demand models because it recognizes the interactions among a series of individual trips, and can also account for connections between trips made by different household members (Niemeier 2005).

Advanced activity-based model systems are now in place, under development, or in planning in more than a dozen large U.S. metropolitan areas and a few medium-sized areas. Figure 16 shows the linkages in lineage among 10 activity-based model systems that are either in operation or well under way in development. Several additional activity-based model systems have been started since this figure was first published, and there is strong reason to believe that activity-based models will become the norm, at least for agencies with significant transportation planning requirements.

In his reviews of the two "families" of activity-based models, Bowman (2009a) noted that all of the model systems have common elements. Figure 17 shows the generic structure that the U.S. activity-based models share. The model systems all—

- Represent an entire day of activities and travel for each member of a synthetic population, using stochastic microsimulation;
- Consist of an integrated system of econometric models; and
- Include traditional traffic and public transport assignment components.

With regard to the PUMS data usage, the key point is that all the activity-based model systems rely on microsimulation-based model application, and therefore also rely on synthetic population modules.

Integrated Land Use Models

Another modeling strategy that transportation planners are using to enhance their analysis capabilities is the development of integrated land use models. Since the 1960s, researchers and planners have been investigating land use models. These efforts have become of significant interest to the transportation planning community.

There is growing recognition that the land use/transportation interaction is significant and must be understood, analyzed, and accounted for to ensure that land use and transportation plans and policies are effective. Most important, there is a growing appreciation of the idea that transportation and land use policies cannot succeed independently of one another (Miller et al. 1999).

Simulating land use improves the outcome of the transportation model. By explicitly simulating the land use and transport interactions, observed behavior of traveling, household relocation, job change, shopping location choice, and the like may be modeled more realistically. The simulation also creates a logical consistency between land use and transportation forecasts, and the performance measures derived from them (Donnelly et al. 2010).

Planners have developed, applied, and continue to use many different integrated land use models. *TCRP Report 48* notes that as of 1999 there were at least 15 separate, implemented land use model applications (Miller et al. 1999). Iacono et al. (2008) have since reviewed the status of land use modeling, and review several land use modeling approaches. As Figure 18 shows, recent integrated land use models have relied on a wide range of conceptual frameworks and specific designs. There has been much recent interest in microsimulation and agent-based modeling in which individual agents (e.g., people, firms) are simulated over time. As for the activity-based travel demand models with microsimulation, the land use model microsimulation formulation has increased the significance of PUMS usage.

Travel Data Simulation

A third type of analysis that is a factor in transportation planners' increased usage of PUMS data is the increased interest and research into survey data transferability.

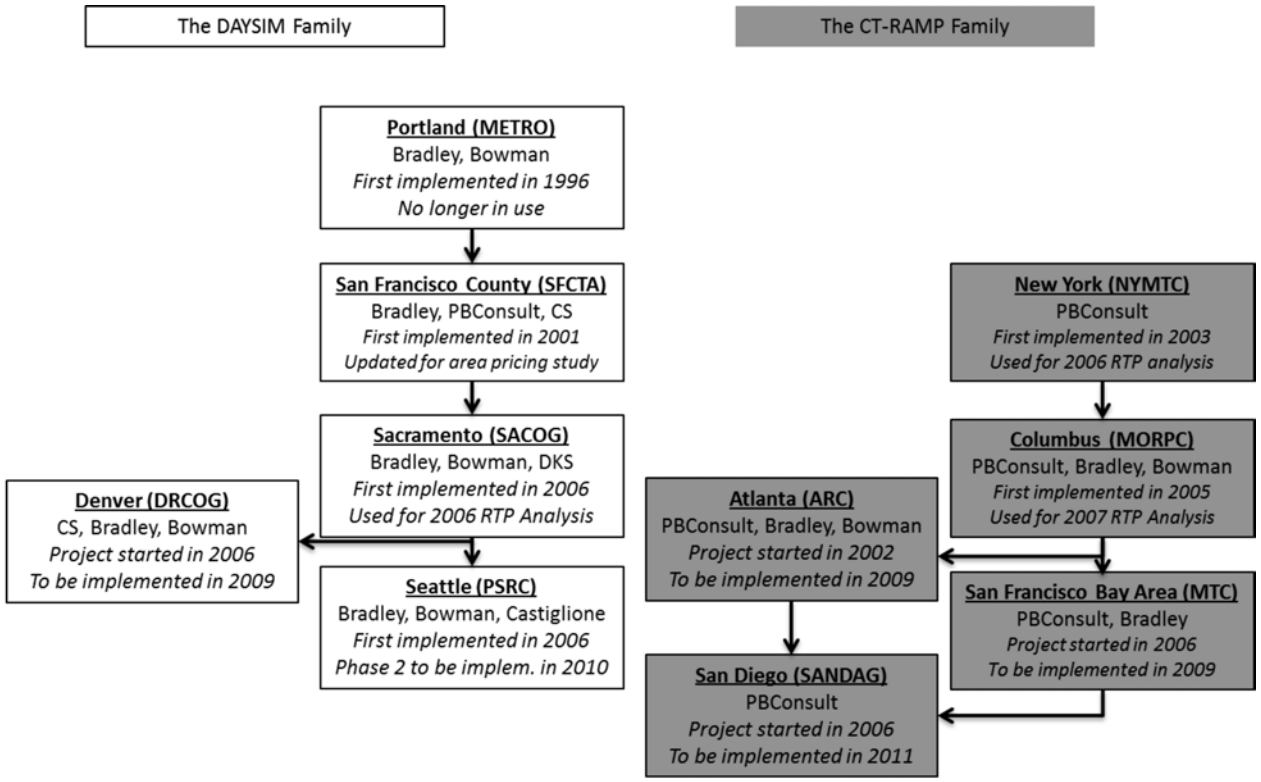


FIGURE 16 Development history of U.S. activity-based model systems. *Source*: Mark Bradley Research and Consulting and J.L. Bowman (2009).

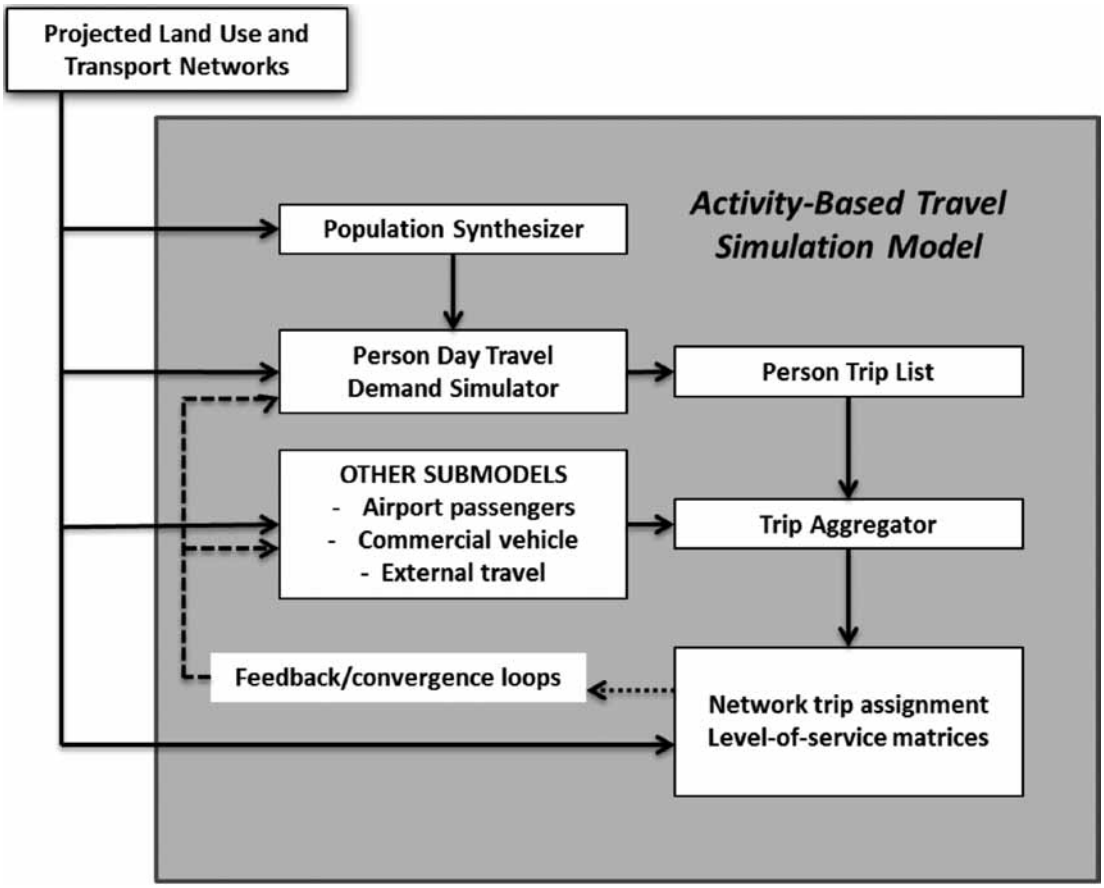


FIGURE 17 General structure of regional activity-based travel demand model systems. *Source*: Mark Bradley Research and Consulting and J.L. Bowman (2009).

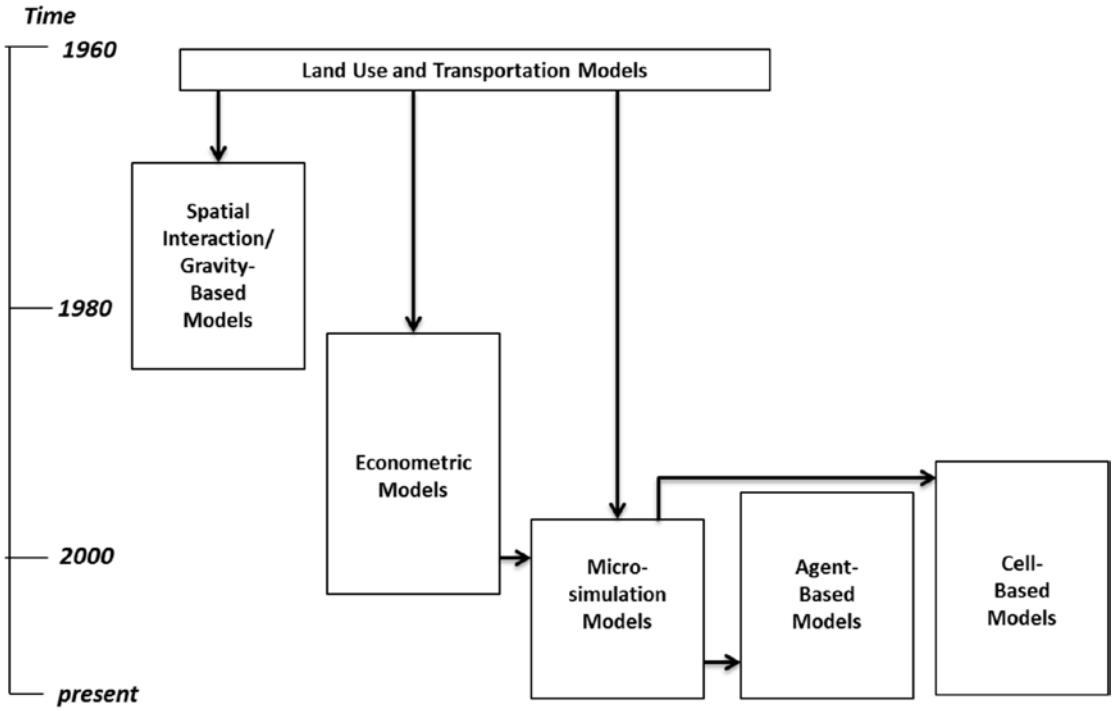


FIGURE 18 Chronology of land use modeling approaches. *Source*: Iacono et al. (2008).

As household travel survey data collection has become more difficult and expensive, transportation researchers have become more interested in developing simulated or model-based survey data. A goal of this research has been to develop ways to combine local socio-demographic data for individuals/households (from sources such as Census Bureau data) with probability distributions of activity/travel patterns (from other travel surveys, such as NHTS) to simulate local travel survey data (Volpe National Transportation Systems Center 2004). Figure 19 shows the general approach for travel survey data simulation summarized by Stopher et al. (2005).

To demonstrate the synthetic data simulation approach, Stopher et al. (2005) used year 2000 Census PUMS data expanded with the Census-provided weights to develop representations of the households within several different project study areas. At the same time, travel behavior data from the NHTS and its forerunner survey, the NPTS, including data on trip rates by purpose, trip lengths, modes, and departure times, were tabulated with key household characteristics, such as household size, vehicle availability, household workers, and presence of children. The authors summarized the travel behavior data into probability distributions by household characteristic combinations. They then used Monte Carlo simulation techniques to assign specific travel behavior characteristics to the individual PUMS records with matching household characteristics.

Based on their initial review of model applications in several metropolitan areas, the authors suggest that the data simulation could be improved through the use of a Bayesian Updating procedure that relies on the incorporation of data from a small local survey. Introducing even a small amount of local data improves the simulated travel behavior data set substantially and cost-effectively (Stopher et al. 2005).

In a study based on New York MSA data, Zhang and Mohammadian (2007) confirmed that the introduction of Bayesian Updating and other innovations involving cluster analyses and neural networking can improve the data simulation results significantly. Even as the simulation techniques improve, the most promising data simulation approaches still rely on the development of synthetic populations, and then the use of Monte Carlo simulation to attach values of travel variables from the survey data.

POPULATION MICROSIMULATION

Although the three analysis techniques described previously seek to introduce different types of improvements to travel demand modeling, they all have come to rely on the application of microsimulation techniques using PUMS data.

Strictly speaking, the broad goals of activity-based models, land use models, and data simulation models could be met

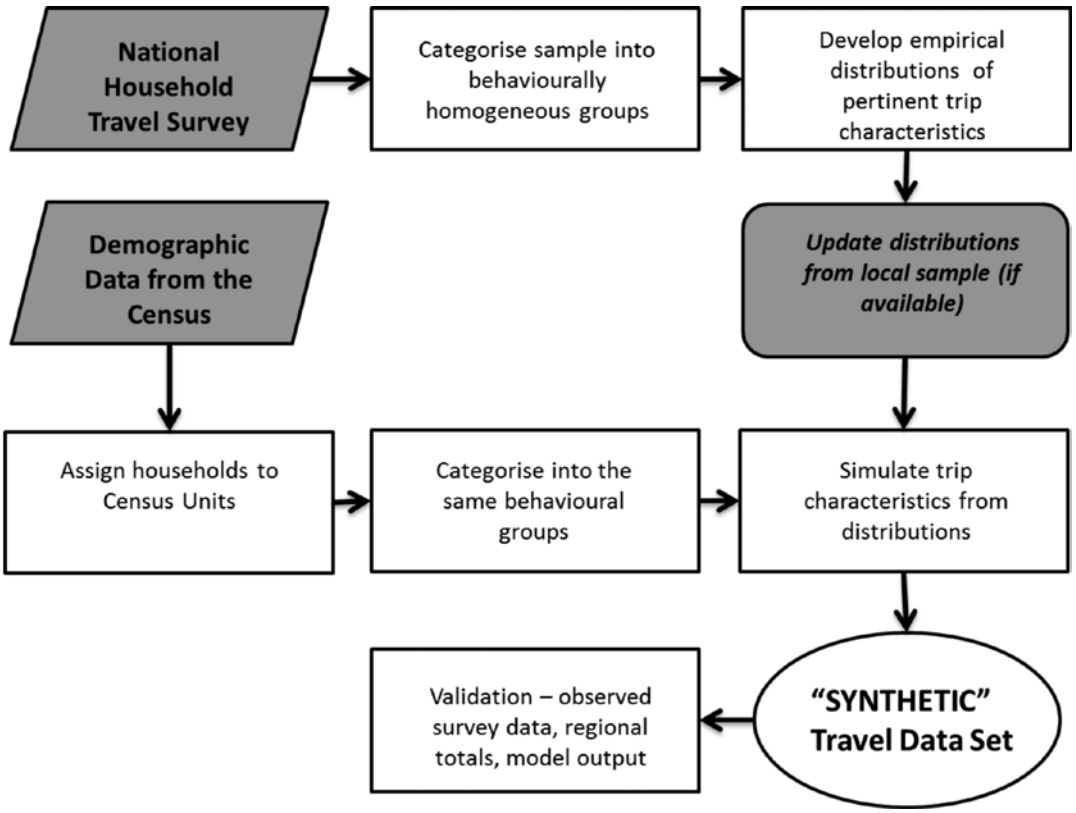


FIGURE 19 General procedure for synthetic travel survey data development. *Source*: Stopher et al. (2005).

without the application of microsimulation techniques. Trip-based models and many types of traditional land use models are applied through the calculation of fractional probabilities to aggregate segments of households at each step of a model system. An important design decision that developers of the advanced models have made is whether the models will be aggregate in nature or whether they will rely on microsimulation. Aggregate models cluster individual model agents (such as households or firms) into homogenous groups that are then analyzed. Microsimulation models analyze the agents as specific units. For each of the three model types, the microsimulation approach has some conceptual advantages over the aggregate approach, as well as computing efficiencies.

Microsimulation models allow complex data sets to be stored more efficiently. Often, microscopic approaches are easier to communicate, as describing the behavior of single actors is less abstract than describing the homogenous behavior of groups. Because microscopic models simulate individual interactions explicitly, model results tend to be more coherent with urban theory (Donnelly et al. 2010).

Because of these advantages, activity-based models are applied through the use of household-specific microsimulation techniques. These models synthesize a set of persons and households that are distributed based on the socioeconomic and demographic characteristics of the study area. Using a population synthesizer permits the model to build consistent marginal distributions of a much wider range of population characteristics. Population synthesis also allows the model to propagate (and reaggregate) these characteristics at later stages in the model.

Travel demand model application using household-by-household simulation had been implemented at least as far back as the Bay Area Short-range Transportation Evaluation Program (STEP) model (Harvey 1978, Ruiter and Ben-Akiva 1978). This model system was trip-based but employed simulation for model application. This modeling innovation did not spread to general practice, and no simulation-based model systems were developed for a number of years.

In the mid-1990s, researchers at Los Alamos National Laboratories developed a travel simulation model, TRANSIMS. The full simulation modeling approach for travel models has not been adopted by transportation planning agencies to any large degree, but in subsequent activity-based modeling efforts, transportation modelers noted the advantages of the TRANSIMS IPF-based household/population synthesis approach. The TRANSIMS population synthesis approach is described by Hobeika (2005). The TRANSIMS population synthesizer and other modeling components are now made available through an open-source agreement, and are supported by an online user community (http://code.google.com/p/transims/). As discussed here, the original TRANSIMS implementations have also led to the development of several other population synthesis approaches.

A new trip-based model that is applied using microsimulation (STEP2) has been developed for Southern Nevada (Walker 2005). As noted earlier, it is theoretically possible to implement an activity-based model using fractional probabilities. Nevertheless, all of the activity-based model systems that have been developed in the United States rely on microsimulation with population synthesis. Similarly, all state-of-the-art land use models and synthetic data models appear to rely on microsimulation with population synthesis.

Population Synthesis Challenge

The goal of population synthesis is to generate a list of households and persons along with many of their detailed housing and population characteristics for each small area zone within the model study area. The challenge is to maintain individual realistic households with all their characteristics, but to assign these households to the small area zones in a way that preserves the zone's demographic distributions for some subset of key household and person characteristics.

Maintaining each household data record with that household's full list of characteristics enables consistency between the microsimulation model components and also allows for the possibility of more explanatory variables in these model components. The microsimulation model applies the model system for each household and person record, thus limiting the introduction of aggregation biases.

Ensuring that base-year demographic distributions of key characteristics are maintained allows the model system to better incorporate forecast changes in these characteristics during model application. If full Census data records with detailed geocoding were available for analysts to use, the data could be used directly in the microsimulation models. However, to maintain confidentiality of Census responses, these data records are not made available. Instead, the Census Bureau provides—

- The PUMS data, which comprise only a sample of Census records and are geocoded only to large geographic areas (PUMAs); and
- Data tabulations with small area estimates of the distribution of categorical household and/or population variables.

Population Synthesis Approaches

Two general approaches for population synthesis to support microsimulation modeling have been proposed: Iterative Proportional Fitting—Synthetic Reconstruction (IPF-SR) and Combinatorial Optimization (CO) (Müller and Axhausen 2011). All of the U.S. transportation agency population synthesizers have been variants of the first approach, so the remain-

ing discussion focuses on that approach. A brief discussion of the CO approach is included at the end of this section.

Bowman (2004) summarizes the general IPF-SR procedure as having two main steps:

1. A multidimensional demographic distribution of households is estimated for each small area zone (e.g., TAZ, census block group, census tract), and

2. A matching sample of households is drawn from a set of household records for which nearly complete Census information is available and is assigned to the small area zone.

The IPF-SR population synthesis process begins with the development of small area joint distributions of a collection of categorical variables that analysts have designated as "control" variables. Decennial Census summary files, ACS multiyear detailed tables, and CTPP multiyear tables provide or will provide categorical variables for small area geographic delineations. For instance, these data sources will provide estimates of the number of households with one, two, three, four, five, six, and seven or more people for a particular block group. The same data sources provide similar information for the number of households by five income categories and the number of households with zero and one or more workers in the block group.

In the first step of the IPF-SR process, analysts develop estimates of the joint distribution of the control variables. So, in the example, we would want to be able obtain estimates of the number of households in each of the 7×5×2=70 household size/income/worker categories. The Census Bureau and CTPP provide several premade cross-tabulations between variables of potential interest, meaning that for many variable combinations the joint distributions are directly available from the Census/AASHTO data sources. The CTPP cross-tabulation for size, income, and workers can be used to obtain estimates for the 70 categories fairly directly through the use of the AASHTO CTPP Data Extraction Tool.

As analysts add more control variables, ready cross-tabulations will no longer be available, so it becomes necessary to use mathematical modeling techniques to develop the joint distribution estimates for the full combination of all control variables. The most common approach, and the one first applied to travel demand model population synthesis by Beckman and the TRANSIMS team, is to apply an IPF procedure (Beckman et al. 1996).

In IPF, analysts assemble a set of control variable marginal totals and an initial joint distribution matrix of control variables. The joint distribution matrix elements are factored to be consistent with one of the control variable marginal totals, and then the adjusted matrix is factored to be consistent with other control variable marginal totals. That output matrix is refactored against the first control variable marginal totals, and the process is repeated until the joint distribution matrix is consistent with all of the control variable marginal totals or a maximum number of iterations are reached (Deming and Stefan 1940). If the control totals are mutually consistent, then IPF eventually converges so that all control totals are satisfied and the correlation structure of the initial joint distribution is preserved (Bowman 2009b).

As noted earlier, for the IPF-SR population synthesis process, most travel demand modelers in the United States have relied on Decennial Census summary files and CTPP tables for the small area marginal totals. Most modelers have relied on processed PUMS data for the initial joint distribution matrices.

The general procedure for developing joint distributions of control variables is schematically described in Figures 20 and 21. Modelers assemble marginal totals for the control variables for TAZs and the PUMS data for the part of the region in which the TAZs are located. The PUMS data are then tabulated by control variable categories at the PUMA geographic level, and these PUMA-level joint distribution matrices are used as the initial joint distributions for the TAZ-level IPF procedures. At the conclusion of the IPF step, joint distributions of control variables for each TAZ have been developed. These distributions maintain the correlation structure of the PUMS data (at the PUMA geographic level), but also the marginal totals for the control variables at the TAZ level.

In the second step of the IPF-SR approach, household records are drawn randomly from the PUMS database and assigned to small areas. This is shown schematically in Figure 22. The assignment of household records is performed through the following steps (Bowman 2009b):

- The IPF process generally results in fractional estimates of households of each type, so the first step of the assignment process is often to round or otherwise render as integers the household estimates.
- Monte Carlo procedures are used to select the correct number of households from the PUMS data set.
- The full set of household and person variables related to the assigned household record are retained for the model application.
- For some synthesis efforts, the households assigned to each small area are assigned to even more detailed geographic areas, such as census blocks or parcels.

The population synthesis procedure is limited by the Census Bureau requirements regarding the size of PUMAs and by the PUMS sampling. If PUMAs were allowed to be smaller, the allocation of households to smaller areas could be obviated or at least made more efficient. In addition, because the PUMS sampling can lead to inconsistencies between PUMS and marginal estimates obtained from

FIGURE 20 Generic approach for developing control variable joint distributions: Data inputs.

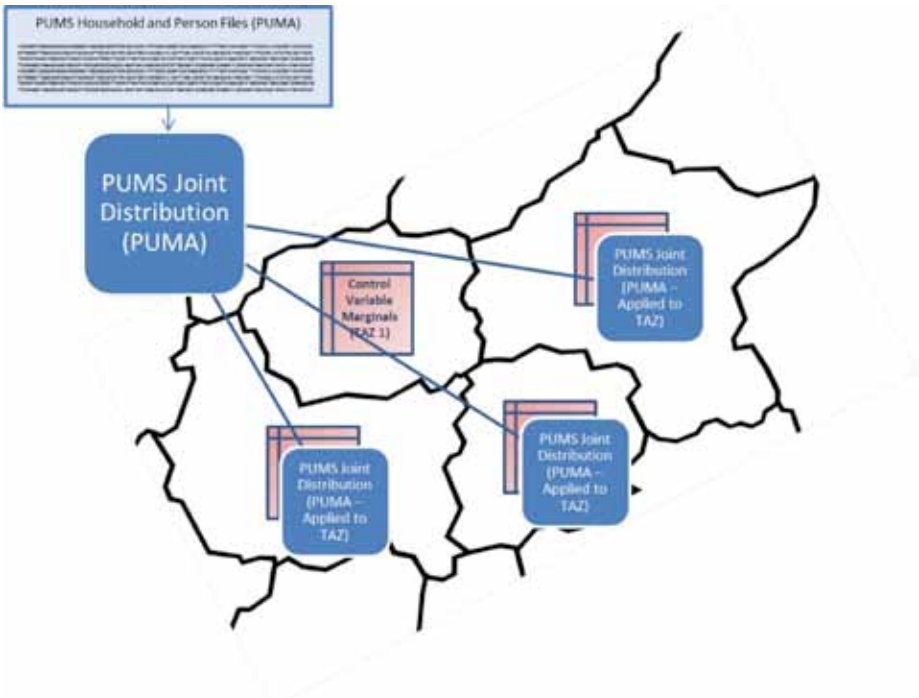


FIGURE 21 Generic approach for developing control variable joint distributions: IPF step.

other Census sources, synthesizer software needs to include a significant amount of computer code to address potential issues. One planner suggested that the Census Bureau consider the disclosure ramifications of providing PUMS data for a greater percentage of ACS records, or even all ACS records (J. Nutting, personal communication, Apr. 2011).

Population Synthesizers

Several different population synthesis efforts that follow the generic overall approach described earlier have been implemented in recent years. Many of the synthesizers have been documented by Bowman (2004, 2009b) and by Müller and

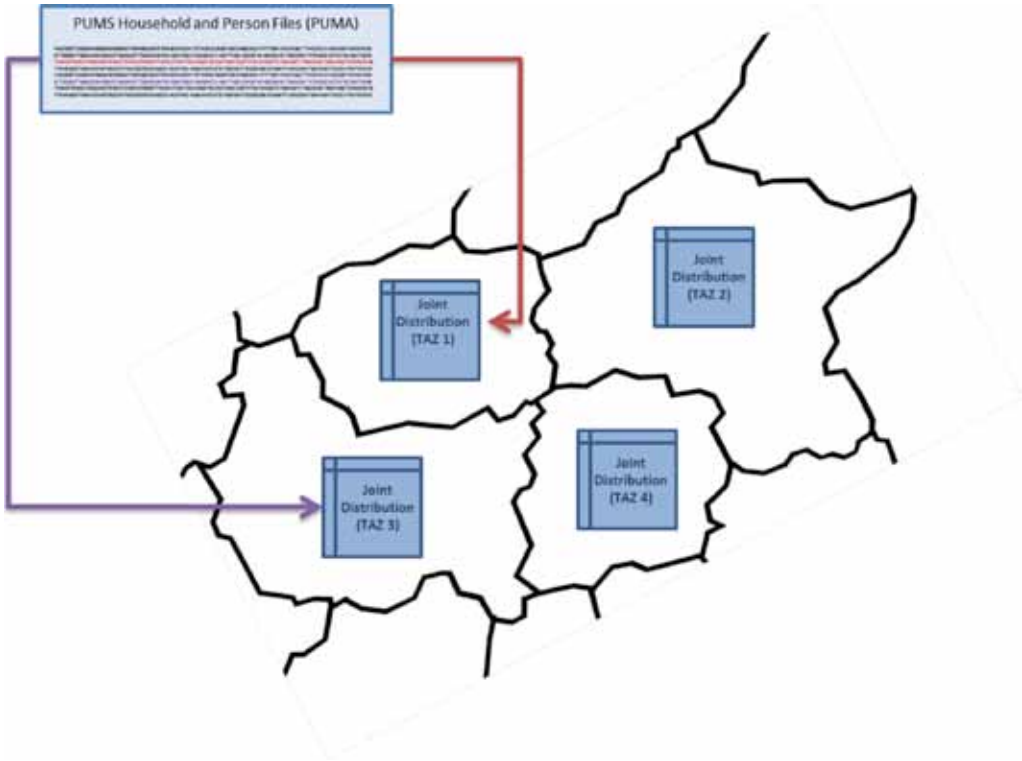


FIGURE 22 Generic approach for assigning Census PUMS records to small areas.

Axhausen (2011). In addition, the synthesizers are described by members of their design teams in the transportation planning literature (see Table 20 for these citations).

Almost all the synthesizers were first developed to address a specific region's modeling needs, but several have been designed (or redesigned) to be applicable to different geographic areas. Transportation planners and land use modelers have developed and improved upon these synthetic population methods over time, and some of the synthesizers have now been coded into full software packages. For instance, the sample generation software created for Atlanta, ARC PopSyn, is designed to provide an extremely flexible system for designating and combining control variables. The software includes facilities for testing how well the synthetic population matches other variables that have not been explicitly controlled. The San Francisco County Transportation Authority, Denver Regional Council of Governments, MTC, and Puget Sound Regional Council model systems are all using derivatives of ARC PopSyn. Similarly, PopGen is now being used in several locations. The PopGen research team is offering to implement, apply, and run PopGen for MPOs, state DOTs, and other agencies interested in generating synthetic populations through open-source licensing arrangements.

Several variations of the IPF-SR approach have developed as planners focus on improving the synthesis process along certain dimensions. Some of the differences in approaches are summarized in Appendix B, and the recent review documents and specific model documentation provide more details (Bowman 2004, 2009b; Müller and Axhausen 2011).

The PopSynWin synthesis approach enables users to consider multilevel controls, including both household and person variables. The synthesizer has been operationalized with a software package (Figure 23) that allows users to select the geography and up to nine control variables for which data are available. The software applies the synthesis routine and provides users with highly visual output statistics and measures of fit.

The PopGen synthesizer employs a heuristic approach, called the Iterative Proportional Updating algorithm, which generates synthetic populations so that both household- and person-level characteristics of interest can be matched in a computationally efficient manner. The PopGen synthesizer algorithm iteratively adjusts weights among the households represented by each cell in the joint distribution until both household- and person-level attributes are matched.

According to the development team, PopGen is being integrated with UrbanSim to offer a seamless ability to simulate population attributes and location choices. The team is now working on the next generation of PopGen, which will incorporate full population evolution and socioeconomic dynamics to evolve the population over time.

Other custom models are being designed with a focus on forecasting incremental population changes over time. They

TABLE 20
POPULATION SYNTHESIZERS IN USE TODAY

Synthesizer	Applications	Population Synthesis Documentation
	TRANSIMS	
Original Deployment	Portland	(Beckman et al. 1996) (Hobeika 2005)
Open Source Deployments	Case Study Applications: - Chittenden County - Atlanta - Buffalo - Sacramento - Portland - Phoenix	http://code.google.com/p/transims/wiki/CaseStudies (RSG 2010)
	Custom Models	
Portland METRO	Oregon statewide (new application)	Under development (Yee 2011)
HGAC	Houston area	Under development (Messen 2011)
SACOG	Sacramento area	(Bowman and Mark Bradley Research and Consulting 2006)
NYMTC	New York metro area	(Parsons Brinckerhoff et al. 2005)
MORPC	Columbus area	(PB Consult 2003)
SEMCOG	Detroit area	Under development (Nutting 2011)
	ARC PopSyn Models	
ARC	Atlanta area	(Bowman and Rousseau 2006)
SFCTA	San Francisco County	(PB and Mark Bradley Research and Consulting 2007)
DRCOG	Denver area	(DRCOG 2011)
PSRC	Seattle/Tacoma area	(Bradley, Bowman, and Castiglione 2008)
MTC	San Francisco Bay Area	Under development (Ory 2011)
	CEMDAP	
NCTCOG	In testing in the Dallas/Fort Worth area	(Bhat et al. 2006) (Guo and Bhat 2007) (Pinjari et al. 2006)
	PopSynWin	
CMAP	In testing in the Chicago area	(Auld 2008) (Auld 2010)
	PopGen /UrbanSim	
MAG	In testing in the Phoenix area	(Ye et al. 2009) http://urbanmodel.asu.edu/popgen.html
FDOT District 7	Tampa area	Under development (Castiglione 2011)
SHRP/FDOT	Jacksonville area	Under development (Castiglione 2011)
Fresno COG	Fresno area	Under development (Castiglione 2011)
San Joaquin Valley MPO	San Joaquin/Stanislaus/ Merced	Under development (Castiglione 2011)
SHRP/CCMPO	Chittenden County	Under development (Castiglione 2011)
	FAMOS	
Florida	In testing in Florida metro areas	(Srinivasan and Ma 2009) (Srinivasan et al. 2008)
	PECAS	
California	Under development in California	(ULTRANS and HBA Specto 2010)

are using the PUMS data to establish the base-year synthetic population and to set the probabilities of certain events such as births, migration, marriage, and divorce. The synthesized population is then modified on a year-by-year basis to develop synthesized population forecasts. The annual releases of ACS PUMS data are useful for this longitudinal approach, but year-to-year changes in the ACS and PUMS data make the analyses more difficult (D. Messen, personal communication, July 2011; D. Yee, personal communication, Apr. 2011).

FIGURE 23 The "PopSynWin" application program. *Source*: Mohammadian (2010).

Müller and Axhausen (2011) raise some concern about the development of alternative synthesizers, rather than a single best one:

> Given the difficulties that routinely arise when trying to properly create a synthetic population, it seems worthwhile to invest time to develop a generic software solution. The software should be applicable to different kinds of input data—concerning both geographic contexts and agent types—without code level changes. Due to the diversity of the input and output data, however, it is likely that a single standalone program will not be able to provide a solution for each possible application. Instead, an extendable open-source software framework that offers routines for tasks that frequently arise in population synthesis applications could be the method of choice (Müller and Axhausen 2011).

The list of population synthesizers in Table 20 does not include those that have been developed solely for research purposes. Nor does the list include any of the many synthesizers that have been developed in other countries, as they are not using PUMS data. It is interesting to note that much of the recent population synthesis research based on overseas applications discusses ways to overcome incomplete or incorrect input data sets. The U.S. research has tended to accept the presence and quality of the PUMS data without any concern.

The different synthesizers can be differentiated by how they address specific analytical issues that are inherent to the basic approach and by design decisions in the population synthesis process. These design parameter decisions are outlined in Appendix B.

Population Synthesis Using the Combinatorial Optimization Approach

Some researchers have proposed an alternative to the general IPF-SR population synthesis approach (Voas and Williamson 2000, Ryan et al. 2009). This approach differs from the synthetic reconstruction methods by eliminating the entire first step of developing small area distribution tables. Instead, as described by Ryan et al. (2009), a randomly selected subset of individuals from the sample is selected, matching the population size of the small area

zone. Statistics are calculated to measure the fit of the subset to the known marginal distributions of control variables in the zone. Then, one of the individuals from the subset is switched with another individual from the sample (with replacement), and the statistics are calculated again. If the overall fit of the new subset is superior to that of the original subset, then the switch is made; otherwise, the original subset is maintained. This process is repeated until threshold values of the comparison statistics are reached, or until a user-defined iteration limit is reached.

The key to success of this straightforward population synthesis approach to problems of any size is the implementation of an efficient optimization algorithm. Many different optimization procedures, including hill-climbing, simulated annealing, and genetic algorithms, could be used to fit the selected sample records to the marginal values of the control variables to more quickly. Lee and Fu (2011) proposed the use of the cross-entropy optimization model, and initially applied the approach to simulate a population of about 10,000 households for Singapore. The authors note that their solution algorithm has many appeals, including the ability to simultaneously address household and population control variables.

Ryan et al. (2009) also demonstrated some success with a CO approach on a smaller-scale synthesis of firms. For this application, the CO approach was found to provide a superior fit to the actual population of firms compared with the IPF-SR approach. Future research will be performed on expanding the CO approach for more sophisticated variable combinations and for larger populations, so future population synthesizers may rely on this approach.

Even if a CO approach were to be used in future U.S. population synthesis efforts, it is likely that PUMS data would still be essential for providing the sample households that would be assigned and reassigned to the small area zones based on the fit with the small area marginal totals of control variables.

SUMMARY OF PUBLIC USE MICRODATA DATA USES

Table 21 summarizes the reasons that the planners that conducted the previously described analyses with PUMS data used these data, and describes the benefits and drawbacks that they identified. As noted earlier, PUMS data offer several unique benefits for users, including the ability to provide the following:

- **Cross-tabulations of variables not readily available from Census or CTPP** – Census and CTPP tables often enable transportation planners to easily locate information needed to support planning applications, but on occasion, analysis needs arise that require combining population characteristics that are not included in the available tabulations. Often, these analyses look at special subpopulations (e.g., members of ethnic groups, people of certain ancestries, group quarters residents) that can be separated using the PUMS data.
- **Cross-tabulations of variables in CTPP but with more currency** – Because PUMS data are available on an ongoing basis and the CTPP are available only periodically, planners can use the PUMS data to create more up-to-date CTPP-like data tables, albeit with less precision in the estimates and less geographic detail.
- **Disaggregate analyses** – Planners and modelers frequently require household- or person-level (disaggregate) data to develop models of the interrelationships between household and person characteristics. The microdata represented by PUMS allow users to evaluate variable relationships at the housing unit and person levels.
- **Comparisons of different regions** – Because the PUMS data series provides common data sets for all regions of the country, and the Census Bureau provides the same attention to detail in its data collection efforts, PUMS data are particularly useful for interregional comparisons and national analyses.
- **Comparisons over time** – PUMS data sets exist for each of the Decennial Census data collection efforts and for each ACS implementation year, so the data are commonly used to track changes in housing and person characteristics over time and changes in the interrelationships between these characteristics over time. Minor changes in the variables and reporting levels make these comparisons more difficult, but planners and researchers are only beginning to explore the utility of annual PUMS data.
- **Validation of other data sources** – PUMS data can be used to independently check calculations and predictions made using other data sources, such as travel surveys, demographic estimates, and modeling results.

The PUMS data uses were also found to be limited in some ways by the nature of the data, including the following points:

- **Lack of geographic specificity** – PUMA definitions are too large for some analyses to be conducted. In many cases where PUMAs were used, there were concerns that important small area geographic variations were glossed over. In addition, PUMA boundaries may differ from study area boundaries, requiring analysts to make judgmental adjustments.
- **Limitations on Census variables** – Several planners indicated that their analyses could have been improved if certain other variables were available in PUMS. Because the PUMS data represent practically the entire set of data from Census microdata records, these analysts were voicing a desire for both an expansion in transportation-relevant Census data, as well as the inclusion in PUMS records of constructed contextual variables that could be used to better understand specific households' relationships with the transportation network.

TABLE 21
SUMMARY OF PUMS DATA USES SUMMARIZED IN THIS SYNTHESIS

PUMS Data Usage	Reasons for Using PUMS Data/Benefits of PUMS Noted by Planners	PUMS Drawbacks Noted by Planners
Houston-Galveston Area Transportation Profile (Ju 2007)	Comparison of PUMS variables over time.	Need for consideration of sampling error in PUMS.
Florida Transportation Profile (Zhou 2004)	Comparison of PUMS variables over time. ACS PUMS found to be consistent with Decennial Census PUMS.	
Housing Custom Tabulations (Griffiths 2011; Purvis 2011)	Development of tabulations which were not available from other Census data sources. Took advantage of the wide range of PUMS variables that were available.	
Environmental Justice and Limited English Proficiency Custom Tabulations (FHWA 2002; Purvis 2011)	Development of tabulations which were not available from other Census data sources. PUMS found to be valuable for detail on occupation and industry, and income by type. PUMS enabled analyses of how household characteristics overlap.	
Gender Custom Tabulations (Krizek et. al. 2004; Weinberger 2007)	Development of tabulations which were not available from other Census data sources. Variable recoding and combination enabled the development of new planning measures. Validation and comparison of other data sources.	Limited by Census having only commute-to-work trip data, and not having trips by other purposes as well.
Immigration Custom Tabulations (Blumenberg et al.; Myers 1996; McGuckin and Srinivasan 2003; Purvis 2003)	Development of tabulations which were not available from other Census data sources, including immigration details not generally available in other datasets. Comparison of PUMS variables over time.	Limited by lack of geographic detail of residences. Limited by Census having only commute-to-work trip data, and not having trips by other purposes as well. Comparisons over time were cross-sectional comparisons, rather than longitudinal comparisons of the same people over time.
Jobs Access Custom Tabulations (UWM 2004; Thakuriah et al. 2005; Hu and Giuliano 2011)	Development of comparisons of regions. Took advantage of the wide range of PUMS household characteristics variables that were available.	Concerns that PUMA geography may be too large to capture some of the jobs access considerations.
Land Use Tabulations (Haas et al. 2006; Duncan 2008; Deal et al. 2009)	Development of comparisons of regions. Comparison of PUMS variables over time. Variable recoding and combination.	Concerns that PUMA geography may be too large to capture some of the land use/transportation interactions. Concerns that Census has only commute-to-work trip data, and not trips by other purposes as well.
Travel Survey Sample Planning (Faussett 2006; Web-scan 2011)	PUMAs used as districts to ensure adequate geographic representation in statewide survey. Took advantage of the wide range of PUMS household characteristics variables that were available. Used PUMS to separate group quarters residents from household residents.	Concerns that PUMA geography may be too large to capture some important travel behavior differences.
Travel Survey Weighting (Nilufar 2003; Gao et al. 2008; Konduri et al. 2009; CS 2011; Payne 2011)	Analyses of data which were more current than data from other Census tabulation products. Took advantage of the wide range of PUMS household characteristics variables that were available. Variable recoding and combination.	PUMA boundaries not the same as study area boundaries. Weighted PUMS estimates may be different than estimates obtained from other Census data sources due to sampling for PUMS.
Travel Survey Weighting (CMAP 2009)	Use of PUMA geography to support analyses.	PUMA boundaries not the same as study area boundaries.
Travel Survey Validation (Pearson et al. 2009; Griffiths 2011)	Analyses of data which were more current than data from other Census tabulation products. Took advantage of the wide range of PUMS household characteristics variables that were available.	Concerns that PUMA geography could be too large to capture some important travel behavior differences in some regions, but not a significant problem for this particular data use.
Travel Demand Modeling for small areas (Ellis 2011)		PUMA delineations too coarse to support modeling in small areas where the number of PUMAs are small.
Travel Demand Modeling: household composition models (Horowitz 2006; Englund et al. 2010; Purvis 2011; Yee 2011)	Development of model components for which detailed geographic delineation is not necessary. Availability of disaggregate (household and person-level) data for modeling. Took advantage of the wide range of PUMS household characteristics variables that were available. Variable recoding and combination.	Concerns that PUMA geography may be too large to capture some important travel behavior differences.

Table 21 continued on p.54

Table 21 continued from p.53

PUMS Data Usage	Reasons for Using PUMS Data/Benefits of PUMS Noted by Planners	PUMS Drawbacks Noted by Planners
Vehicle Availability Models (CS 1996, 1997; Ryan and Han 1999; Baber 2004; Purvis 2004;)	Development of model components for which detailed geographic delineation is not necessary. Availability of disaggregate (household and person-level) data for modeling. Validation and comparison of other data sources	Concerns that PUMA geography may be too large to capture some important travel behavior differences.
Internal-External Modeling (Griffiths 2011)	Took advantage of availability of consistent PUMS data nationwide. Availability of disaggregate (household and person-level) data for modeling.	
Travel Demand Model Validation (Purvis 2011; Web-scan 2011)	Validation and comparison of other data sources. Variable recoding and combination.	Concerns that PUMA geography may be too large to capture some important travel behavior differences.
Land use/Transportation Modeling (TELUS 2005; Weinberger and Goetzke 2009; Morris and Smart 2011)	Availability of disaggregate (household and person-level) data for modeling. Analyses of data which were more current than data from other Census tabulation products. Took advantage of the wide range of PUMS household characteristics variables that were available.	Concerns that PUMA geography may be too large to capture some important travel behavior differences. Concerns that Census has only commute-to-work trip data, and not trips by other purposes as well.
Market Segmentation Modeling (Beckman et al. 2008; Zhou et al. 2004)	Availability of disaggregate (household and person-level) data for modeling.	Concerns that PUMA geography may be too large to capture some important travel behavior differences. Concerns that Census has only commute-to-work trip data, and not trips by other purposes as well.
Modeling of Immigrant Transportation (Cline et al. 2009; Kim 2009; Chatman and Klein, 2011)	Availability of disaggregate (household and person-level) data for modeling. Took advantage of the wide range of PUMS household characteristics variables that were available. Variable recoding and combination.	Additional data variables, such as driver's license status, could be used to improve PUMS analyses. Concerns that PUMA geography may be too large to capture some important travel behavior differences, but PUMA geography was found to be adequate for some spatial analyses.
Population Microsimulation Models (Various)	Availability of disaggregate (household and person-level) data for modeling. Took advantage of the wide range of PUMS household characteristics variables that were available. Analyses of data which were more current than data from other Census tabulation products.	Concerns that PUMA geography may be too large to capture some important travel behavior differences. Weighted PUMS estimates may be different than estimates obtained from other Census data sources due to sampling for PUMS. Changes in PUMS categories from year to year require coding modifications and extra work for analysts.

- **Year-to-year inconsistencies in PUMS files** – The ACS has changed some questions since its introduction. These changes are reflected in the PUMS data files, so data users must match data dictionaries to analyze successive years. Multiyear PUMS files reflect the most recent data dictionaries, but analytical processes and code have often been developed with the older variable definitions. The PUMS data files are not backward compatible.
- **Sampling error for PUMS (especially for ACS PUMS)** –The ACS and the former Decennial Census long form data are samples drawn from the population, with estimates that contain sampling error. The PUMS data represent samples of those samples, so it is essential that planners understand the confidence limits on PUMS-based analyses and realize that PUMS estimates can vary from other estimates.

CHAPTER FIVE

CONCLUSIONS AND FURTHER RESEARCH

This synthesis effort has sought to examine why and how transportation planners use the Census Public Use Microdata Sample (PUMS) data series. The answers to these questions appear to depend on which of the two groups of transportation agency data users being examined.

One group of agencies considers PUMS data important elements in the array of data that are essential for fulfilling the agencies' technical missions. For this group, the PUMS data support advanced travel demand modeling efforts and allow users to analyze population subgroups for which data are not always easy to find. PUMS data enable planners to be innovative, and to design custom research that would otherwise require expensive new primary data collection activities. This group is generally happy with the PUMS data, but would like to see several specific areas of technical improvements addressed or at least studied, especially as they move forward and will need to rely solely on American Community Survey (ACS)-based PUMS data.

Another group of transportation agencies does not clearly understand the PUMS data. These agency planners are unfamiliar or only vaguely familiar with the PUMS data, and they have not yet seen evidence that an investment of time and resources in learning more will provide a direct positive return on their investment. This group would benefit from a better understanding of advantages, disadvantages, and potential uses of the PUMS data so that they can make more informed decisions about how to use the PUMS data. Although there are frequent PUMS data users among both larger metropolitan planning organizations and state departments of transportation and smaller agencies, the more frequent users tend to be the larger agencies that have more technical resources and that have jurisdiction over several different Public Use Microdata Areas (PUMAs). Many smaller agencies have jurisdiction over areas that are represented by only one or two PUMAs.

Further evaluation and research into the following activities would help address the needs of these two groups identified in the synthesis effort:

Research on ways to improve knowledge sharing between the Census Bureau and the transportation planning community regarding the PUMS data series. Through the Census Transportation Planning Products (CTPP) project, the transportation community has come to a better understanding of Census issues, and the Census Bureau has developed a better understanding of transportation planning data uses. Training and outreach appear to be needed in the incorporation of sampling error concepts into analyses. PUMS data users need guidance in understanding the margins of error in estimates and analyses that are based on PUMS data, and they need to know effective ways to present their findings in ways that reflect these margins of error.

Conduct research to determine the feasibility of improvements in the PUMS data series. A number of PUMS data users interviewed as part of this synthesis effort expressed interest in learning more about the rationale for some of the limiting aspects of the PUMS data series. Generally speaking, and perhaps surprisingly, the PUMS data users supported robust data disclosure avoidance. Many expressed the wish that PUMAs could be drawn with smaller minimum populations, although they appeared to understand and support the reasons for the population requirement.

Some analysts believed that it would be helpful to have research conducted on PUMS limitations such as the 100,000 minimum population requirement to determine whether there could be alternative approaches for future data releases. Data users would like to see the following specific issues analyzed in further research:

- Whether the 1 percent sampling rate for ACS PUMS could be enlarged without damaging disclosure risk levels significantly;
- Whether more specific journey-to-work related estimates could be provided without damaging disclosure risk levels significantly; and
- Whether the minimum PUMA population of 100,000 could be changed without damaging disclosure risk levels significantly.

In addition, the Census Bureau has begun to develop a microdata access system, an online table generator that can be used to conduct tabulations and statistical analyses of Census microdata while maintaining confidentiality. Such a system may in the future allow for PUMS-like tabulations based on the full ACS data set, rather than on a sample of the ACS data. Transportation planning could benefit from the this type of system, provided that it is able to provide the necessary analyses.

Conduct research on the best ways to improve knowledge sharing among the transportation planning community regarding the PUMS data series. Transportation planning agencies, and especially agencies with fewer technical staff, would benefit from access to other planners' PUMS-based studies, as well as their data manipulation software procedures, data processing code, and tools. The synthesis literature review identified opportunities for practitioners to collaborate, such as on multiregional analyses, and opportunities to share ideas, such as when data users have common research topics. However, this synthesis is by no means a complete accounting of PUMS data uses by transportation planners. Because the Census PUMS data are often used in support of larger transportation planning analyses, the PUMS usage often is not formally documented. Instead, the PUMS data analyses are not documented or are documented only in internal "gray literature" (e.g., memos, technical notes, internal emails). A centralized location for the documentation of PUMS data usage and for sharing PUMS resources, such as a wiki-based resource, could be established either under the auspices of an existing institution, such as a TRB committee or Integrated Public Use Microdata Series, or through a new social media–based user forum.

Promote continuing research and dissemination of research on issues related to population synthesis. Research on population synthesis methods is ongoing and active. Among the key issues that could be analyzed further are developments in constrained optimization synthesis procedures, selection of effective control variables, and further work on heuristics to address known issues with the current synthesis approaches. The emergence of stand-alone software packages for population synthesis may enable researchers to evaluate and compare alternative specifications more easily, and to gain a better understanding of how best to validate synthesis results. In addition, the application of the same population synthesizers in multiple locations may enable researchers to better discriminate between generalizable and site-specific research conclusions.

Promote new research on population synthesis design decisions in light of the Census Bureau's migration to ACS and the introduction of the new synthetic CTPP tables. New research is needed on the following questions:

- What is the effect of using multiyear PUMS data, rather than data from a single year, for the development of iterative proportional fitting joint distributions?
- What is the effect of using multiyear PUMS data, rather than data from a single year, for household allocation to small zones?
- What is the effect of relying on the soon-to-be-released CTPP tables that have been perturbed to avoid disclosure issues for control variable marginal totals, and should this change in CTPP tables affect the selection of population synthesis control totals? (This research is relevant to the use of PUMS data, because the synthesized tables may introduce greater levels of inconsistency between the PUMS inputs and the control total inputs.)
- Is it technically and institutionally feasible (and effective in terms of output quality) for population syntheses to be performed on larger microdata samples within the Census Bureau, and then for the resulting simulation results to be disclosure-proofed and released?

GLOSSARY

The following glossary entries were taken from Census documentation and other online sources. (U.S. Census Bureau 2009a, McNally 2010).

American Community Survey (ACS). An ongoing statistical survey by the U.S. Census Bureau, sent to approximately 250,000 addresses monthly (or 3 million per year). It regularly gathers information previously contained only in the long form of the Decennial Census. It is the largest survey other than the Decennial Census that the Census Bureau administers.

American FactFinder (AFF). An electronic system for access to and dissemination of Census Bureau data on the Internet. AFF offers prepackaged data products and user-selected data tables and maps from Census 2000, the 1990 Census of Population and Housing, the 1997 and 2002 Economic Censuses, the Population Estimates Program, annual economic surveys, and the ACS.

Block group. A subdivision of a census tract (or, prior to 2000, a block numbering area), a block group is a cluster of blocks having the same first digit of their four-digit identifying number within a census tract.

Calibration. The procedure used to adjust travel models to simulate base-year travel.

Census geography. A collective term referring to the types of geographic areas used by the Census Bureau in its data collection and tabulation operations, including their structure, designations, and relationships to one another. See http://www.census.gov/geo/www/index.html.

Census tract. A small, relatively permanent statistical subdivision of a county delineated by a local committee of census data users for the purpose of presenting data. Census tract boundaries normally follow visible features, but may follow governmental unit boundaries and other nonvisible features; they always nest within counties. Designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions at the time of establishment, census tracts average about 4,000 inhabitants.

Confidence interval. The sample estimate and its standard error permit the construction of a confidence interval that represents the degree of uncertainty about the estimate. A 90% confidence interval can be interpreted roughly as providing 90% certainty that the interval defined by the upper and lower bounds contains the true value of the characteristic.

Confidentiality. The guarantee made by law (Title 13, U.S. Code) to individuals who provide census information, regarding nondisclosure of that information to others.

Consumer Price Index (CPI). The CPI program of the Bureau of Labor Statistics produces monthly data on changes in the prices paid by urban consumers for a representative basket of goods and services.

Controlled. During the ACS weighting process, the intercensal population and housing estimates are used as survey controls. Weights are adjusted so that ACS estimates conform to these controls.

Current Population Survey (CPS). A monthly survey of about 50,000 households conducted by the Census Bureau for the Bureau of Labor Statistics. The CPS is the primary source of information on the labor force characteristics of the U.S. population.

Current residence. The concept used in the ACS to determine who should be considered a resident of a sample address. Everyone who is currently living or staying at a sample address is considered a resident of that address, except people staying there for 2 months or less. People who have established residence at the sample unit and are away for only a short period of time are also considered to be current residents.

Custom tabulations. The Census Bureau offers a wide variety of general purpose data products from the ACS. These products are designed to meet the needs of the majority of data users and contain predefined sets of data for standard census geographic areas, including both political and statistical geography. These products are available on the American FactFinder and ACS websites. For users with data needs not met through the general purpose products, the Census Bureau offers "custom" tabulations on a cost-reimbursable basis, with the ACS Custom Tabulation program. Custom tabulations are created by tabulating data from ACS microdata files. They vary in size, complexity, and cost depending on the needs of the sponsoring client.

Decennial Census. Data collection mandated by the U.S. Constitution. The population is enumerated every 10 years and the results are used to allocate congressional seats (congressional apportionment), electoral votes, and government program funding. The year 2000 Decennial Census (and several previous ones) included the collection of long form data from a sample of participants. The long form data collection has been supplanted by the ACS.

Disaggregate demand model. Model that is obtained by using the observations of the travel choice behavior of individuals directly for model calibration and that is usually probabilistic.

Disclosure avoidance (DA). Statistical methods used in the tabulation of data prior to releasing data products to ensure the confidentiality of responses. See **Confidentiality**.

Distribution. Process by which trips defined by origin are distributed among the various available destinations. Common trip distribution models are the gravity model and the opportunity model.

District. A grouping of contiguous zones that are aggregated to larger areas.

Dwelling unit. A room or group of rooms, occupied or intended for occupancy as separate living quarters, by a family or other group of persons living together or by a person living alone.

Estimates. Numerical values obtained from a statistical sample and assigned to a population parameter. Data produced from the ACS interviews are collected from samples of housing units. These data are used to produce estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology.

Five-year estimates. Estimates based on 5 years of ACS data. These estimates reflect the characteristics of a geographic area over the entire 5-year period and will be published for all geographic areas down to the census block group level.

Fratar distribution. A method of distributing trip ends based on the growth factor of the origin and destination and on the given trip interchanges. Named for Thomas J. Fratar.

Generation. Step in the sequential, aggregate forecasting process in which trips defined by origin or destination (but not both) are predicted based on the characteristics of the activity system and, in some applications, some measure of transportation service to or from the zone. The output of generation is a one-dimensional array of trips into or out of a zone for input to trip distribution models.

Geographic summary level. A geographic summary level specifies the content and the hierarchical relationships of the geographic elements that are required to tabulate and summarize data. For example, the county summary level specifies the state-county hierarchy. Thus, both the state code and the county code are required to uniquely identify a county in the United States or Puerto Rico.

Group quarters (GQ) facility. A place where people live or stay that is normally owned or managed by an entity or organization providing housing and/or services for the residents. These services may include custodial or medical care, as well as other types of assistance. Residency is commonly restricted to those receiving these services. People living in GQ facilities are usually not related to each other. The ACS collects data from people living in both housing units and GQ facilities.

Group quarters (GQ) population. The number of persons residing in GQ facilities.

Integrated Public Use Microdata Series (IPUMS). Microdata samples from United States (IPUMS-USA) and international (IPUMS-International) Census records. The records are converted into a consistent format and made available to researchers through a web-based data dissemination system. IPUMS is housed at the Minnesota Population Center, an interdisciplinary research center at the University of Minnesota, under the direction of Professor Steven Ruggles.

Item allocation rates. A method of imputation used when values for missing or inconsistent items cannot be derived from the existing response record. In these cases, the imputation must be based on other techniques such as using answers from other people in the household, other responding housing units, or people believed to have similar characteristics. Such donors are reflected in a table referred to as an allocation matrix. The rate is the percentage of times this method is used.

Iterative proportional fitting (IPF). Mathematical procedure (also known as bi-proportional fitting in statistics, RAS algorithm in economics, and matrix raking or matrix scaling in computer science) is an iterative algorithm for estimating cell values of a contingency table such that the marginal totals remain fixed and the estimated table decomposes into an outer product.

Logit model. Analytical form for demand modeling that is suited to modeling of multiple travel choice situations.

Margin of error (MOE). The difference between an estimate and its upper or lower confidence bounds. Confidence bounds can be created by adding the MOE to the estimate (for the upper bound) and subtracting the MOE from the estimate (for the lower bound). All published ACS MOE are based on a 90-percent confidence level. Some ACS products provide an MOE instead of confidence intervals.

Mode of travel. Means of travel such as auto driver, vehicle passenger, mass transit passenger, or walking.

Model. A mathematical formula that expresses the actions and interactions of the elements of a system in such a manner that the system may be evaluated under any given set of conditions (e.g., land use, economic, socioeconomic, travel characteristics).

Multiple regression. Sometimes used interchangeably with multiple correlation, but normally used with reference to the regression equation resulting from correlation analysis.

Multiyear estimates. Three- and 5-year estimates based on multiple years of ACS data. Three-year estimates will be published for geographic areas with a population of 20,000 or more. Five-year estimates will be published for all geographic areas down to the census block group level.

National Household Travel Survey (NHTS). A survey that provides information to assist transportation planners and policymakers who need comprehensive data on travel and transportation patterns in the United States. The 2009 NHTS updates information gathered in the 2001

NHTS and in prior Nationwide Personal Transportation Surveys conducted in 1969, 1977, 1983, 1990, and 1995.

Nonsampling error. Total survey error can be classified into two categories: sampling error and nonsampling error. Nonsampling error includes measurement errors due to interviewers, respondents, instruments, and mode; nonresponse error; coverage error; and processing error.

North American Industry Classification System (NAICS). Classification system used by business and government to classify business establishments according to type of economic activity (process of production) in North America. It was designed to replace the older Standard Industrial Classification (SIC) system.

Period estimates. An estimate based on information collected over a period of time. For ACS, the period is 1 year, 3 years, or 5 years.

Point-in-time estimates. An estimate based on one point in time. The Decennial Census long form estimates for Census 2000 were based on information collected as of April 1, 2000.

Population Estimates Program. Official Census Bureau estimates of the population of the United States, states, metropolitan areas, cities and towns, and counties; also official Census Bureau estimates of housing units (HUs).

Public Use Microdata Area (PUMA). An area that defines the extent of territory for which the Census Bureau releases Public Use Microdata Sample (PUMS) records.

Public Use Microdata Sample (PUMS) files. Computerized files that contain a sample of individual records, with identifying information removed, showing the population and housing characteristics of the units, and people included on those forms.

Puerto Rico Community Survey (PRCS). The counterpart to the ACS that is conducted in Puerto Rico.

Reference period. Time interval to which survey responses refer. For example, many ACS questions refer to the day of the interview; others refer to "the past 12 months" or "last week."

Residence rules. The series of rules that define who (if anyone) is considered to be a resident of a sample address for purposes of the survey or census.

Sampling error. Errors that occur because only part of the population is directly contacted. With any sample, differences are likely to exist between the characteristics of the sampled population and the larger group from which the sample was chosen.

Sampling variability. Variation that occurs by chance because a sample is surveyed rather than the entire population.

Simulation. To reproduce synthetically; for example, to simulate a trip distribution.

Single-year estimates. Estimates based on the set of ACS interviews conducted from January through December of a given calendar year. These estimates are published each year for geographic areas with a population of 65,000 or more.

Standard error. A measure of the deviation of a sample estimate from the average of all possible samples.

Statistical significance. The determination of whether the difference between two estimates is not likely to be from random chance (sampling error) alone. This determination is based on both the estimates themselves and their standard errors. For ACS data, two estimates are "significantly different at the 90 percent level" if their difference is large enough to infer that there was a less than 10% chance that the difference came entirely from random variation.

Subarea, subregion. Normally, an analysis area that is significantly smaller than the usual metropolitan region and is important because many alternatives influence only subareas.

Three-year estimates. Estimates based on 3 years of ACS data. These estimates are meant to reflect the characteristics of a geographic area over the entire 3-year period. These estimates will be published for geographic areas with a population of 20,000 or more.

Traffic analysis zone (TAZ). Unit of geography most commonly used in conventional transportation planning models. The size of a zone varies, but for a typical metropolitan planning software, a zone of under 3,000 people is common. The spatial extent of zones typically varies in models, ranging from large exurban areas to spaces as small as city blocks or buildings in central business districts.

Trip assignment. The process of determining route or routes of travel and allocating the zone-to-zone trips to these routes.

Trip distribution. The process by which the movement of trips between zones is estimated. The data for each distribution may be measured or be estimated by a growth factor process, or by synthetic model.

Trip generation. A general term describing the analysis and application of the relationships which exist between the trip makers, the urban area, and the trip making. It relates to the number of trip ends in any part of the urban area,

Trip length frequency distribution. The array that relates the trips or the percentage of trips made at intervals or various trip distances.

Trip purpose. The reason for making a trip, normally one of ten possible purposes. Each trip may have a purpose at each end; for example, home to work.

Trip table. A table showing trips between zones, either directionally or total two-way. The trips may be separated by mode, purpose, time period, vehicle type, or other classification.

REFERENCES

Auld, J., A.K. Mohammadian, and K. Wies, "Population Synthesis with Control Category Optimization," presented at the 10th International Conference on Application, 2008.

Auld, J., A.K. Mohammadian, and K. Wies, "An Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels," presented at the 89th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 10–14, 2010.

Azimi, E., "The PUMS and IPUMS of 2000 Census," presented at the Transportation Research Board Conference on Census Data for Transportation Planning: Preparing for the Future, Irvine, Calif., May 11–13, 2005.

Baber, C., "Baltimore Region Model Application Auto Availability Estimation Using 2001," presented at Data for Understanding Our Nation's Travel National Household Travel Survey Conference, Nov. 1–2, 2004.

Barton–Aschman Associates, Inc., and Cambridge Systematics, Inc., *Model Validation and Reasonableness Checking Manual,* Travel Model Improvement Program, FHWA, Washington, D.C., Feb. 1997.

Beckman, J.D., et al., "Immigration, Residential Location, Car Ownership, and Commuting Behavior: A Multivariate Latent Class Analysis from California," presented at the 87th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 13–17, 2008.

Beckman, R.J., K.A. Baggerly, and M.D. McKay, "Creating Synthetic Baseline Populations," *Transportation Research Part A: Policy and Practice*, Vol. 30, No. 6, 1996, pp. 415–429.

Bhat, C.R., et al., *The Comprehensive Econometric Microsimulator for Daily Activity-travel Patterns (CEMDAP)*, Report 4080-S, prepared for the Texas Department of Transportation, Austin, 2006.

Blumenberg, E., and Evans, A.E. (2010) "Planning for Demographic Diversity: The Case of Immigrants and Public Transit," *Journal of Public Transportation*, Vol. 13, No. 2, 2010, pp. 23-45.

Blumenberg, E. and K. Shiki, "Transportation Assimilation: Immigrants, Race and Ethnicity, and Mode Choice," presented at the 86th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 21–25, 2007.

Blumenberg, E. and K. Shiki, "Immigrants and Resource Sharing: The Case of Carpooling," presented at the 87th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 13–17, 2008.

Blumenberg, E. and L. Song, "Travel Behavior of Immigrants in California: Trends and Policy Implications," presented at the 87th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 13–17, 2008.

Bowman, J.L., "A Comparison of Population Synthesizers Used in Microsimulation Models of Activity and Travel Demand," 2004 [Online]. Available: http://jbowman.net/papers/2004.Bowman_Comparison_of_PopSyns.pdf [accessed July 29, 2010].

Bowman, J.L., "Historical Development of Activity Based Model Theory and Practice," *Traffic Engineering and Control*, Vol. 50, No. 2, pp. 59–62 (part 1); Vol. 50, No. 7, pp. 314–318 (part 2), 2009a.

Bowman, J.L., "Population Synthesizers," *Traffic Engineering and Control*, Vol. 49, No. 9, 2009b, p. 342.

Bowman, J. and Mark Bradley Research and Consulting, SACSIM/05: Activity-Based Travel Forecasting Model for SACOG Featuring *DAYSIM*—the Person Day Activity and Travel Simulator: Technical Memo Number 2 Population Synthesis, July 31, 2006, Draft 4.

Bowman, J.L. and G. Rousseau, Validation of the Atlanta (ARC) Population Synthesizer (PopSyn), white paper presented at the TRB Conference on Innovations in Travel Modeling, Austin, Tex., May 21–23, 2006.

Bradley, M., J. Bowman, and J. Castiglione, *Activity Model Work Plan &Activity Generation Model: Work Plan Report,* prepared for Puget Sound Regional Council, Sep. 29, 2008.

Cambridge Systematics, Inc., *New Hampshire Statewide Planning Study Vehicle Availability Model*, Cambridge Systematics, Inc., June 24, 1996.

Cambridge Systematics, Inc., *Enhancement of DVRPC's Travel Simulation Models: Task 10, Vehicle Availability Model*, Philadelphia, Pa., Apr. 1997.

Cambridge Systematics, Inc., Cincinnati GPS Household Survey Project Memorandum on Survey Weighting, May 2011.

Cambridge Systematics, Inc., NuStats, N. McGuckin, and E. Ruiter, *NCHRP Report 588: A Guidebook for Using American Community Survey Data for Transportation Planning,* Transportation Research Board of the National Academies, Washington, D.C., 2007.

Chatman, D.G. and N. Klein, "Immigrants and Automobility in New Jersey: The Role of Spatial and Occupational Factors in Commuting to Work," presented at the 90th

Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 23–27, 2011.

Chicago Metropolitan Agency for Planning (CMAP), "Weighting the Chicago Regional Household Travel Inventory Survey With 2005–2007 American Community Survey Data and an Eleven Zone Geographic System," Dec. 2009 [Online]. Available: http://www.cmap.illinois.gov/travel-tracker-survey.

Cline, M., C. Sparks, and K. Eschbach, "Understanding Carpool Use Among Hispanics in Texas," *Proceedings of the 88th Annual Meeting of the Transportation Research Board* (DVD), Washington, D.C., Jan. 11–15, 2009.

Deal, B., J.H. Kim, and V.G. Pallathucheril, "Growth Management and Sustainable Transport: Do Growth Management Policies Promote Transit Use?" P*roceedings of the 88th Annual Meeting of the Transportation Research Board* (DVD), Washington, D.C., Jan. 11–15, 2009.

Deming, W.E. and F.F. Stephan, "On the Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known," *Annals of Mathematical Statistics*, Vol. 11, No. 4, 1940, pp. 427–444.

Denver Regional Council of Governments (DRCOG), "Focus Model Overview," 2011 [Online]. Available: http://www.drcog.org/index.cfm?page=FocusTechnicalResources.

Donnelly, R., G. Erhardt, R. Moeckel, and W. Davidson, *NCHRP Synthesis 406: Advanced Practices in Travel Forecasting*, Transportation Research Board of the National Academies, Washington, D.C., 2010.

Duncan, M., "Comparing Rail Transit Capitalization Benefits for Single-Family and Condominium Units in San Diego, California," *Transportation Research Record: Journal of the Transportation Research Board, 2067*, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp 120–130.

Englund, D., R. Eash, and M. Lupa, "Matching Workers and Employment Opportunities: Linking Employees and Workplaces by Earnings in Regional Travel Models," Presentation to the Transport Chicago 2010 Conference, Chicago, Ill., June 4, 2010.

Faussett, K.M., "MDOT's Statewide Household Travel Survey—Design, Implementation, and Lessons Learned," presented at the 85th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 22–26, 2006.

Federal Highway Administration (FHWA), *Transportation and Environmental Justice Effective Practices*, Report FHWA-EP-02-016, FHWA, Washington, D.C., Jan. 2002.

Gao, R., B. (Brenda) Zhou, and K.M. Kockelman, "Opportunities for and Impacts of Carsharing: A Survey of the Austin, Texas Market," presented at the 87th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 13–17, 2008.

Guo, J.Y. and C.R. Bhat, "Population Synthesis for Microsimulating Travel Behavior," *Transportation Research Record, Journal of the Transportation Research Board, 2014*, Transportation Research Board of the National Academies, Washington, D.C., Vol. 12, 2007, pp. 92–101.

Haas, P., C. Makarewicz, A. Benedict, T. Sanchez, and C. Dawkins, "Housing and Transportation Cost Trade-Offs and Burdens of Working Households in 28 Metros," Center for Neighborhood Technology, Chicago, Ill., July 2006.

Harvey, G., "STEP: Short-range Transportation Evaluation Program," prepared for the Metropolitan Transportation Commission, Oakland, Calif., 1978.

Heither, C., *CMAP Travel Demand Model Validation Report*, Feb. 23, 2011.

Hobeika, A., TRANSIMS Fundamentals: Chapter 3: Population Synthesizer, Technical Report, Virginia Polytechnic University, Blacksburg, Va., July 2005.

Horowitz, A., *NCHRP Synthesis 358: Statewide Travel Forecasting Models*, Transportation Research Board of the National Academies, Washington, D.C., 2006.

Hu, L. and G. Giuliano, "Beyond the Inner City: A New Form of Spatial Mismatch," *Proceedings of the 90th Annual Meeting of the Transportation Research Board* (DVD), Washington, D.C., Jan. 23–27, 2011.

Iacono, M., D. Levinson, and A. El-Geneidy, "Models of Transportation and Land Use Change: A Guide to the Territory," *Journal of Planning Literature Online First*, Feb. 13, 2008.

Ju, S., *Transportation Planning Capacity Building Program, Peer Exchange Report—Using ACS Data in Transportation Planning Applications*, Daytona Beach, Fla., May 10–11, 2007.

Kim, S., "Immigrants and Transportation: An Analysis of Immigrant Workers' Work Trips," *Cityscape: A Journal of Policy Development and Research*, Vol. 11, No. 3, 2009.

Konduri, K., H. Bar-Gera, B. Sana, X. Ye, and R.M. Pendyala, "Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods," presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 11–15, 2009.

Krenzke, T., "NCHRP 08-79: Current Research Efforts to Minimize Effects of Disclosure in the CTPP," *CTPP Status Report*, Aug. 2010 [Online]. Available: http://www.fhwa.dot.gov/ctpp/sr0810.htm.

Krizek, K.J., P.J. Johnson, and N. Tilahun, "Gender Differences in Bicycling Behavior and Facility Preferences," *Conference Proceedings 35 Research on Women's Issues in Transportation*, Nov. 18–20, 2004.

Lee, D.-H. and Fu, Y.F., "A Cross Entropy Optimization Model for Population Synthesis Used in Activity-Based Micro-Simulation Models," presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 23–27, 2011.

Mark Bradley Research and Consulting and J.L. Bowman, "Strategy for Activity-Based Travel Demand Model Development with Travel Survey: Final Report," prepared for Southern California Association of Governments Project 09-012, June 26, 2009.

McGuckin, N. and N. Srinivasan, "National Summary," *Journey to Work Trends in the United States and its Major Metropolitan Areas 1960–2000*, Publication No. FHWA -EP-03-058, 2003.

McNally, M., Travel Forecasting Glossary 2010 [Online]. Available: http://www.its.uci.edu/~mcnally/tdf-glos.html.

McWethy, L., "Analysis of Iterative Proportion Fitting in the Generation of Synthetic Populations," *CTPP Status Report*, April 2009 [Online]. Available: http://www.fhwa.dot.gov/ctpp/sr0409.htm.

Metropolitan Transportation Commission Planning Section, *Bay Area Travel Survey 2000 (BATS2000) Sample Weighting and Expansion: Working Paper #1*, June 2003.

Miller, E., D. Kriger, and J.D. Hunt, *TCRP Report 48: Integrated Urban Models for Simulation of Transit and Land Use Policies: Guidelines for Implementation and Use*, Transportation Research Board of the National Academies, Washington, D.C., 1999.

Mohammadian, A., "Transferability of Travel Survey Data: A Household Travel Data Simulation Tool," presented to Travel Model Improvement Program Webinar, Jan. 25, 2010.

Morris, E.A. and M.J. Smart, "Surgeons and Smog: Expert vs. Lay Perception of the Risks of Automobile-Generated Air Pollution," presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 23–27, 2011.

Müller, K. and K.W. Axhausen, "Population Synthesis for Microsimulation: State of the Art," presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 23–27, 2011.

Murakami, E., *PUMS and PUMAs, CTPP Status Report*, Apr. 2009 [Online]. Available: http://www.fhwa.dot.gov/ctpp/sr0409.htm.

Myers, D., "Changes over Time in Transportation Mode for Journey to Work: Effects of Aging and Immigration," *Decennial Census Data for Transportation Planning: Case Studies and Strategies for 2000*, Vol. 2: Case Studies, Transportation Research Board, Washington, D.C., 1996.

Niemeier, D., "Activity-Based Models and ACS Data: What Are the Implications for Use?" prepared for the Transportation Research Board Conference on Census Data for Transportation Planning: Preparing for the Future, Irvine, Calif., May 11–13, 2005.

Nilufar, F., Assessing Sampling Biases and Establishing Standardized Procedures for Weighting and Expansion of Data, *Proceedings of the 9th TRB Conference on the Application of Transportation Planning Methods*, Baton Rouge, La., Apr. 6–10, 2003.

Parsons Brinckerhoff, PB Consult, AECOM Consult, Urbitran Associates, Urbanomics, Alex Anas & Associates, NuStats International, George Hoyt & Associates, *New York Best Practice Model (NYBPM)*, Final Report 2005 [Online]. Available: http://www.nymtc.org/project/BPM/model/bpm_finalrpt.pdf.

PB, in association with Mark Bradley Research & Consulting, *Update of the San Francisco Chained Activity Modeling Process (SF-CHAMP)*, prepared for the San Francisco County Transportation Authority, June 11, 2007.

PB Consult, *Task 2: Household and Population Synthesis Procedure, PB Consult/Parsons Brinckerhoff*, prepared for the Mid-Ohio Regional Planning Commission as part of the MORPC Model Improvement Project, Mar. 12, 2003.

Pearson, D., et al., "Improving Accuracy in Household and External Travel Surveys," Aug. 2009, published Jan. 2010 [Online]. Available: http://tti.tamu.edu/documents/0-5711-1.pdf.

Pinjari, A.R., et al., *Activity-Based Travel-Demand Analysis for Metropolitan Areas in Texas: CEMDAP Models, Framework, Software Architecture and Application Results*, Research Report, 4080–8, Texas Department of Transportation, Department of Civil, Architectural and Environmental Engineering, University of Texas. Austin, Oct. 2006.

Purvis, C.L., "Using 1990 Census Public Use Microdata Sample to Estimate Demographic and Automobile Ownership Models," *Transportation Research Record 1443*, Transportation Research Board, TRB, National Research Council, Washington, D.C., 1994, pp. 21–29.

Purvis, C., "Commuting Patterns of Immigrants," *CTPP Status Report*, Aug. 2003 [Online]. Available: http://www.fhwa.dot.gov/planning/census_issues/ctpp/status_report/sr0803.pdf.

RSG, Inc., PopSyn (version 4.0), Apr. 20, 2010 [Online]. Available: http://code.google.com/p/transims/wiki/DocumentationIndex.

Ruiter, E. and M. Ben-Akiva, "Disaggregate Travel Demand Models for the San Francisco Area: System Structure, Component Models and Application Procedures," In *Transportation Research Record 673*, Transportation Research Board, National Research Council, Washington, D.C., 1978, pp. 121–128.

Ryan, J., H. Maoh, and P.S. Kanaroglou, "Population Synthesis: Comparing the Major Techniques Using a Small,

Complete Population of Firms," *Geographical Analysis,* Vol. 41, No. 2, 2009, pp. 181–203.

Ryan, J.M. and G. Han, "Vehicle-Ownership Model Using Family Structure and Accessibility Application to Honolulu, Hawaii," *Transportation Research Record 1676*, Transportation Research Board, National Research Council, Washington, D.C., 1999.

Srinivasan, S., L. Ma, and K. Yathindra, *Procedure for Forecasting Household Characteristics for Input to Travel-Demand Models, Final Report*, TRC-FDOT-64011-2008, Transportation Research Center, University of Florida, Tampa, 2008.

Srinivasan, S. and L. Ma, "Synthetic Population Generation: A Heuristic Data-Fitting Approach and Validations," presented at the 12th International Conference on Travel Behaviour Research (IATBR), Jaipur, Rajasthan, India, Dec. 2009.

Stopher, P.R., P.S. Greaves, and M. Xu, "Using National Data to Simulate Metropolitan Area Household Travel Survey," *Journal of Transportation Statistics*, Vol. 8, No. 3, 2005.

TELUS PROJECT, *TELUM (Transportation Economic and Land Use Model) User's Manual, Version 5.0*, Mar. 2005.

Thakuriah (Vonu), P., P.S. Sriraj, P.S., S. Soot, Y. Liao, and G. Berman, "Activity and Travel Changes of Job Access Transportation Service Users: Analysis of a User Survey," *Transportation Research Record: Journal of the Transportation Research Board, 1927*, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 55–62.

Transportation Research Board, *Special Report 288: Metropolitan Travel Forecasting: Current Practice and Future Direction*, Transportation Research Board of the National Academies, Washington, D.C., 2007.

ULTRANS and HBA Specto Incorporated, "CSTDM09—California Statewide Travel Model Development: Population (Draft Final System Documentation)," Sep. 2010 [Online]. Available: http://ultrans.its.ucdavis.edu/files/pecas/CSTDM09_Population_DraftFinal.pdf.

U.S. Census Bureau, *2000 Census of Population and Housing, Public Use Microdata Sample, United States: Technical Documentation*, Washington, D.C., 2003.

U.S. Census Bureau, *A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know*, U.S. Government Printing Office, Washington, D.C., 2008.

U.S. Census Bureau, *A Compass for Understanding and Using American Community Survey Data: What PUMS Data Users Need to Know*, U.S. Government Printing Office, Washington, D.C., 2009a.

U.S. Census Bureau, *A Compass for Understanding and Using American Community Survey Data: What Researchers Need to Know*, U.S. Government Printing Office, Washington, D.C., 2009b.

U.S. Census Bureau Geography Division, *Final Public Use Microdata Area (PUMA) Criteria and Guidelines for the 2010 Census and the American Community Survey*, Aug. 3, 2011 [Online]. Available: http://www.census.gov/geo/puma/2010_puma_guidelines.pdf.

University of Wisconsin–Milwaukee (UWM), Center for Economic Development, "Transportation Equity and Access to Jobs in Metropolitan Milwaukee," UWM Center for Economic Development, Sep. 2004.

Voas, D. and P. Williamson, "An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata," *International Journal of Population Geography*, Vol. 6, 2000, pp. 349–366.

Volpe National Transportation Systems Center, "Summary Report for the Peer Exchange on Data Transferability," Washington, D.C., Dec. 16, 2004.

Vovsha, P., J. Bowman, and M. Bradley, "Activity-Based Travel Forecasting Models in the United States: Progress Since 1995 and Prospects for the Future," In *Progress in Activity-Based Analysis*, H.J.P. Timmermans, Ed., Elsevier Science, Oxford, U.K., 2005, pp. 389–414.

Walker, J.L., "Making Household Microsimulation of Travel and Activities Accessible to Planners," presented at the 84th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 9–13, 2005.

Wargelin, L., Stopher, P., et al., "GPS-Based Household Interview Survey for the Cincinnati, Ohio Region," prepared for Ohio Department of Transportation, Office of Research and Development and the U.S. Department of Transportation, Federal Highway Administration, Feb. 2012 [Online]. Available: http://www.dot.state.oh.us/Divisions/Planning/SPR/Research/reportsandplans/Reports/2012/Planning/134421_FR.pdf

Weinberger, R., "Men, Women, Job Sprawl and Journey to Work in the Philadelphia Region," *Public Works Management and Policy*, Vol. 11, No. 3, Jan. 2007, pp. 177–193.

Weinberger, R. and F. Goetzke, "How Previous Experience Affects Automobile Ownership," *Proceedings of the 88th Annual Meeting of the Transportation Research Board* (DVD), Washington, D.C., Jan. 11–15, 2009.

Ye, X., K. Konduri, R.M. Pendyala, B. Sana, and P. A. Waddell, "A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations," presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 11–15, 2009.

Zhang, Y. and A. Mohammadian, "Investigating the Transferability of National Household Travel Survey Data," *Transportation Research Record: Journal of the Transportation Research Board, No. 1993*, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 67–79.

Zhou, L., "An Analysis of Journey to Work Characteristics in Florida Using Census 2000 Public Use Microdata Sample Data Files," *Theses and Dissertations*, Paper 1317, 2004 [Online]. Available: http://scholarcommons.usf.edu/etd/1317.

Zhou, Y., K. Viswanathan, Y. Popuri, and K.E. Proussaloglou, "Transit Customers—Who, Why, Where, and How: A Market Analysis of the San Mateo County Transit District," presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 2004.

APPENDIX A

Survey Questionnaire

Use and Application of Census PUMS by MPOs and States

Introduction

Dear Transportation Planner:

The Transportation Research Board (TRB) is preparing a synthesis on how transportation planners at the state and local level use federal data sources, in general, and the Census Bureau Public Use Microdata Sample database, in particular. This is being done for NCHRP, under the sponsorship of the American Association of State Highway and Transportation Officials, in cooperation with the Federal Highway Administration.

The web survey will seek to shed light on the breadth of PUMS usage by planners at state departments of transportation and metropolitan planning organizations. The synthesis results will allow planners to see how others use these Census data, and to help identify potential improved transportation planning products using these data.

This survey is being sent to state and regional transportation planners and demographic experts. Your cooperation in completing the questionnaire will ensure the success of this effort. If you are not the appropriate person at your agency to complete this survey, please forward it to the correct person.

Please complete and submit this survey by March 28, 2011. We estimate that it should take no more than 15 minutes to complete. If you have any questions, please contact our principal investigator, Kevin Tierney (kevintierney@rocketmail.com; 617.839.0938). Any supporting materials can be sent directly to Kevin Tierney by e-mail or at the postal address shown at the end of the survey.

QUESTIONNAIRE INSTRUCTIONS

To view and print the entire questionnaire, Click on the following link: //appv3.sgizmo.com/users/64484/Use_and_Application_of_Census_PU.doc and print using "control p."

To save your partial answers, or to forward a partially completed questionnaire to another party, click on the "Save and Continue Later" link in the upper right hand corner of your screen. A link to the partial survey will be e-mailed to you or a colleague.

To view and print your answers before submitting the survey, click forward to the page following question 19. Print using "control p."

To submit the survey, click on "Submit" on the last page.

The questions with an asterisk require a response of some kind.

Please enter the date (MM/DD/YYYY).

_____

Please enter your contact information.

First Name: _____

Last Name: _____

Title: _____

Agency/Organization: _____

Street Address: _____

Suite: _____

City: _____

State: _____

Zip Code: _____

Country: _____

E-mail Address: _____

Phone Number: _____

Fax Number: _____

Mobile Phone: _____

URL: _____

Census Data Use

1.) Which description below best expresses your role within your agency with regard to using U.S. Census Bureau data?

[ ] I am not personally responsible for analyses that use U.S. Census Bureau data

[ ] I am one of a group of staff members that are responsible for analyses that use U.S. Census Bureau data

[ ] I am the only one in the agency that is responsible for analyses that use U.S. Census Bureau data

Alternative Contacts

1a) Who within your agency is responsible for analyses that use U.S. Census Bureau data and might be available to complete this survey?

[ ] No one within our agency performs analyses that use U.S. Census Bureau data

[ ] I am not sure who within our agency performs analyses that use U.S. Census Bureau data

[ ] Contact 1 Name & E-mail:

[ ] Contact 2 Name & E-mail:

[ ] Contact 3 Name & E-mail:

Unqualified Thank/Terminate

Thank you for your input. We will send our survey invitation to the others within your agency you have identified. If you have any questions about this survey, please contact Kevin Tierney at kevintierney@rocketmail.com or 617-839-0938. If you would prefer to have Mr. Tierney contact you directly, please enter your email address or telephone number below.

_____

Additional Contacts for Qualified

2.) Which of the following statements best describes your familiarity with the Census data analyses that others in your agency are responsible for?

[ ] I am not the only one responsible for analyses using U.S. Census data, but I am basically familiar with almost all the uses of these data within our agency

[ ] I am not sure who else within our agency performs analyses that use U.S. Census Bureau data

[ ] Other agency staff members may be able to provide information about our agency's usage of Census Bureau data that I cannot

2a) The following agency staff members may be able to provide information about our agency's usage of Census Bureau data that I cannot

Contact Person 1 Name & E-mail: _____

Contact Person 2 Name & E-mail: _____

Contact Person 3 Name & E-mail: _____

Data Product Familiarity

3.) Please indicate how much you and others in your agency use each of the following data sources by selecting the familiarity level that best describes your usage of each data set.

	Regular User of These Data Source	Occasional User of These Data Source	Familiar with These Data Source But Do Not Use It	Not Familiar with This Data Source
Decennial Census Tabulations	( )	( )	( )	( )
American Community Survey (ACS) Tabulations	( )	( )	( )	( )
Census Transportation Planning Products (CTPP)	( )	( )	( )	( )
Census Longitudinal Employer-Household Dynamics (LEHD)	( )	( )	( )	( )
Census Public Use Microdata Sample (PUMS)	( )	( )	( )	( )
Census Annual Population Estimates	( )	( )	( )	( )
Census Economic Surveys (Annual Survey of Manufactures, County/Zip Code Business Patterns, Nonemployer Statistics)	( )	( )	( )	( )
Bureau of Labor Statistics Employment Databases	( )	( )	( )	( )
FHWA National Household Travel Survey (NHTS)	( )	( )	( )	( )
State or Local Household Travel Survey(s)	( )	( )	( )	( )

Number of Data Users

4.) About how many agency staff members access U.S. Census Bureau data and other federal data sources, such as those listed above, as part of their job responsibilities?

_____

Importance of Decennial Tabulations

5.) In your view, how important are Decennial Census Tabulations to your agency's mission?

[ ] Very important/central to agency's mission

[ ] Somewhat important

[ ] Useful, but not too important

[ ] Not useful

Importance of ACS Tabulations

6.) In your view, how important are American Community Survey Tabulations to your agency's mission?

[ ] Very important/central to agency's mission

[ ] Somewhat important

[ ] Useful, but not too important

[ ] Not useful

Importance of CTPP

7.) In your view, how important are Census Transportation Planning Products to your agency's mission?

[ ] Very important/central to agency's mission

[ ] Somewhat important

[ ] Useful, but not too important

[ ] Not useful

Importance of PUMS

8.) In your view, how important are Census Public Use Microdata Sample data to your agency's mission?

[ ] Very important/central to agency's mission

[ ] Somewhat important

[ ] Useful, but not too important

[ ] Not useful

Reasons for Using PUMS

9.) For which of the following purposes do you use PUMS data?

[ ] Data cross-tabulations

[ ] Weighting and expansion of travel surveys

[ ] Travel demand modeling components

[ ] Synthetic population microsimulation

[ ] Other PUMS analyses

9a) What types of crosstabulations or other PUMS analyses do you perform?

Analysis Type 1: _____

Analysis Type 2: _____

Analysis Type 3: _____

PUMS Based Analyses

10.) We are very interested in finding out about how transportation planners are using the Census PUMS data. Would you be willing to provide us with additional information regarding your use of Census PUMS data?

[ ] Yes

[ ] No

10a) Can you provide the following?

[ ] Written documentation or reports that describe your use of Census PUMS data

[ ] A short written summary (e-mail) that describes your use of Census PUMS data

[ ] Participation in a brief telephone interview about your use of Census PUMS data

[ ] Contact information for others involved in your use of Census PUMS data

Means of Accessing PUMS Data

11.) Do you obtain Census PUMS data…

[ ] Directly from the Census Bureau website or data center?

[ ] From a University or company that disseminates the data in its original format?

[ ] From a University or company that tabulates the data or provides a means for you to tabulate data?

Assessment of PUMS (General)

12.) How would you rate the following aspects of the Census PUMS data?

	Excellent	Good	Fair	Poor	No Opinion/Not Sure
Ease of accessing the PUMS data	( )	( )	( )	( )	( )
Ease of manipulating and analyzing the PUMS data	( )	( )	( )	( )	( )
Availability and quality of the documentation for the PUMS data	( )	( )	( )	( )	( )

13.) What improvements would you like to see made to the Census PUMS data?

Improvement 1: _____

Improvement 2: _____

Improvement 3: _____

Improvement 4: _____

Assessment of PUMS (Data quality)

14.) How would you rate the following aspects of the Census PUMS data quality?

	Excellent	Good	Fair	Poor	No Opinion/Not Sure
PUMS sample sizes	( )	( )	( )	( )	( )
PUMS geographic definitions	( )	( )	( )	( )	( )
PUMS data disclosure avoidance procedures	( )	( )	( )	( )	( )
PUMS commuting and workplace data	( )	( )	( )	( )	( )

14a) What specific problems do the Census PUMS data have? What improvements would you like to see made to the Census PUMS data?

Improvement 1: _____

Improvement 2: _____

Improvement 3: _____

Improvement 4: _____

Comments on PUMS

15.) Do you have any comments on the usefulness of Census PUMS data for the analyses your agency performs?

_____

_____

_____

_____

Reasons for Not Using PUMS

16.) You indicated that your agency does not use Census PUMS data. Which of the following reasons describe why you do not use PUMS data?

[ ] Not completely aware of what the PUMS data are

[ ] No need for PUMS given availability of other data sources

[ ] Lack technical knowledge and/or time needed to use PUMS

[ ] Software and computing limitations

[ ] Not satisfied with PUMS quality and consistency

[ ] Not satisfied with PUMS weighting and sampling

[ ] Not satisfied with PUMS geographic area sizes

[ ] Not satisfied with PUMS workplace location / commuting

[ ] Other reasons for not using PUMS

16a) What improvements would you like to see made to the Census PUMS data before you used this data source?

Improvement 1: _____

Improvement 2: _____

Improvement 3: _____

Improvement 4: _____

Agency Type

17.) The last questions will be used to classify your responses. Which of the following statements describe your agency?

[ ] A state department of transportation

[ ] A metropolitan planning organization for an area with a population of less than 100,000

[ ] A metropolitan planning organization for an area with a population of between 100,000 and 200,000

[ ] A metropolitan planning organization for an area with a population of between 200,000 and 500,000

[ ] A metropolitan planning organization for an area with a population of between 500,000 and 1,000,000

[ ] A metropolitan planning organization for an area with a population of more than 1,000,000

Agency Size

18.) What is the total dollar amount, for all purposes, expressed in your MPO's Unified Planning Work Program?

_____

19.) How many people work for your agency?

Full-time employees: _____

Part-time employees: _____

Thank You!

Thank you for taking our survey. Your response is very important to us. If you have any questions or comments, please feel free to contact Kevin Tierney at:

E-mail: kevintierney@rocketmail.com

Mail: 206 Broad Meadow Road, Needham, MA 02492

Phone: 617.839.0938

APPENDIX B

Population Synthesis Design Issues

The key design issues for the population synthesizers include:

Step 1—Small area joint distributions

- Selection of control variables
- Definition of control variable categories
- Mechanisms for addressing the "zero cell" problem
- Mechanisms for maintaining person level joint distributions, as well as household level joint distributions

Step 2—Assignment of household and person data records to small areas

- Mechanisms for preparing the joint distributions for PUMS record allocation
- Sample draw procedures

Selection of Control Variables and Definition of Control Variable Categories

Bowman provides a summary of the control variables used in the implementations of many of the population synthesizers (Bowman 2009). As he notes, almost all of the model implementations have used zone-level data and forecasts of household size and income as control variables for sampling households from the regional PUMS households. In addition, most of the regions have used the number of workers in the household as a third control variable, both because it is important behaviorally, and because a CTPP table (Table 1-75 in the year 2000 CTPP) provides a useful 3-way joint distribution of household size, number of workers and income for 2000.

The SACOG population synthesizer uses only these three control variables so that it is possible to obtain the necessary small area joint distribution without performing IPF on PUMS data. Other synthesizers have included several other control variables:

- The Portland (METRO) and San Francisco (SFCTA) models have also used age of head of household as a control variable,
- San Francisco (SFCTA) is also using controls for presence of children, single vs. multi-family dwelling, and race/ethnicity. The SFCTA model is explicitly synthesizing residents of group quarters housing.
- Atlanta (ARC), MTC and Denver are all analyzing the use of age or age-related variables (e.g., presence of children and/or senior citizens).
- For a recent application of the ARC PopSyn for the Atlanta Region, the following PUMS data items were used to develop the joint distribution (Rousseau 2011):
- Household type (institutionalized vs. non-institutionalized),
- Age,
- Personal income, and
- Employment status.

Since students in college dorms might rely on family income that does not appear on their census form, it was possible to "promote" them to higher income classification based on other data in the PUMS record (e.g., student and poverty status). The PUMS records of non-institutionalized GQ were likewise included in the 1-person household groups for purposes of drawing PUMS households into the synthetic population (Rousseau 2011).

The initial PopGen application included control variables for both households and people. It had Household Type (Family: Married Couple; Family: Male Householder, No Wife; Family: Female Householder, No Husband; Non-family: Householder Alone; Non-family: Householder Not Alone), Household size, and Household income, as well as person level controls for gender, age, and race.

The initial CEMDAP application included as control variables: Household family status, household type (similar to PopGen), presence of children, presence of retirement age household members, and household size, as well as person level con-

trols for gender, age, and race. As discussed below, these synthesizers were designed to analyze a combination of household and person variables.

PopSynWin was initially tested with various combinations of control variables, including household size, number of workers, income, presence of children, presence of retirement age household members, household type, and vehicles.

The differences in the range of control variables for the different synthesis implementations probably indicate:

1. The model developers have different hypotheses about how household characteristics relate to each other and to the transportation and land use system, and/or

2. There are not yet reliable procedures for evaluating the effectiveness and usefulness of control variable selections.

There is some evidence that the selection of appropriate control variables will affect results. McWethy (2009) compared two-way tables of household income by household size resulting from the performance of an IPF-based population synthesis generation using PUMS data and TAZ level household income and household size control variables against TAZ level two-way tables available directly from CTPP tabulations. Using chi-squared tests, she found the two-way tables were significantly different from one another for more than a quarter of the TAZs.

Based on these results, it is important that analysts apply care to the IPF analyses, and that analysts take advantage of available small area cross-tabulations, such as those provided by CTPP, in establishing population synthesis control variable marginals (McWethy 2009).

Fortunately, the sample generation software created for ARC PopSyn, PopSynWin, and PopGen allow future users to designate and combine control variables. The software packages provide facilities for testing how well the synthetic population matches control variables and other variables that have not been explicitly controlled.

One of the key design features of the PopSynWin software is that it is designed to make the re-categorization of control variables efficient. Users are able to set categories according to their specific needs and computing resources, without having to manipulate the original data sources.

Mechanisms for Addressing the "zero cell" Problem

Because the PUMS data and the marginal totals data from CTPP or Census summary files are based on different samples, they will not always be completely consistent. A problem arises in the IPF application when there is a non-zero marginal for a specific category, but no records in the PUMS data in that category. As Muller and Axhausen (2011) suggest, the most direct solution to this problem is to replace the false zero with a small value, so that the IPF routine does not need to divide by zero. However, this solution introduces bias, so several of the synthesis procedures address the problem in other ways.

The PopSynWin program relies on its efficient re-categorization capabilities to redesign the joint distribution matrix so that the problem cell(s) are collapsed. This method is also recommended by the CEMDAP development team. The ARC PopSyn and FSUTMS synthesizers identify adjacent cells to swap values with. The PopGen synthesizer uses a more sophisticated approach to model a non-zero value based on the overall regional PUMS data.

Mechanisms for Maintaining Person Level Joint Distributions

Generally, when synthetic households are drawn from the PUMS records, they will faithfully reflect the household variable controls. Usually, uncontrolled household variables will also be similar to marginal estimates of those variables. On the other hand, the person records in the households that are drawn from the PUMS data may or may not compare well with marginal estimates.

The developers of the CEMDAP and PopGen synthesizers sought to apply more sophisticated means than standard IPF to address this problem. Guo and Bhat (CEMDAP) developed an algorithm that generates both household and person level joint distributions, and allows one to relax the household fit in order to improve the person-level fit (Guo and Bhat 2007).

Ye et al. demonstrate a new heuristic approach to the factoring problem, which they have dubbed iterative proportional updating (IPU), that develops joint distributions that match well both on household and person-level control variables. The

algorithm performs IPF like factoring, but introduces household and person record weights that get adjusted and recalculated during the factoring until both household and person attributes are matched (Ye et al. 2009).

Mechanisms for Assigning Household and Person Data Records to Small Areas

The population synthesizers also vary in terms of the details on how they draw PUMS records to assign to small areas. As noted above, the first step of this process is usually to "integerize" the joint distribution matrices. Then, household and person records are drawn from the PUMS for each geographic area. The synthesizers use a wide variety of procedures for both of these steps, as outlined by Bowman (2009). In many cases, the rounding of the joint distribution cell values introduces bias, but some of the synthesizers are designed to minimize or control the direction of the bias.

The population synthesizers also differ in terms of the algorithms that they use to draw household and person records from the PUMS file to assign to individual small area zones. Some of the synthesizers draw households randomly, and then test whether the drawn record still "fits" within the small area zone to determine whether it is accepted or rejected. Some of the synthesizers have weighting and sorting algorithms that are used in an effort to make the assignment process be more efficient. Some draw with replacement; others draw without replacement. Finally, some synthesizers will include a sample of household data records from adjacent PUMAs in the pool of records from which households are drawn. Having these additional records improves the ability of the synthesizer to match the multidimensional targets more easily.

Once again, there is wide variation in the synthesizer practices, but little empirical evidence to show whether one approach is better than another. As alternative strategies are employed, modelers will develop better understanding of the costs and benefits of different approaches.

Need for Re-application of the Synthesizer

The final difference between population synthesis implementations is that for some applications (primarily for land use modeling), the entire process needs to be re-performed in order to assign the households and person records for the small area zones to more detailed geography, such as the block or parcel level.

Synthesizer Design Issues Pertaining to the Use of PUMS

For almost all of the IPF/SR synthesizers, the Census PUMS data provide two essential inputs:

- Joint distribution seed matrix for the IPF/IPU process; and
- The sample list from which household and population records are drawn to be assigned to small area zones.

For the first input, there is an open research question as to whether it makes sense to apply the unadjusted PUMA level joint distribution tables to the smaller areas. The assumption is that the correlation structure should be maintained across the geographic levels, but it may be a testable hypothesis.

A second concern with the IPF procedure going forward will be the suitability of using synthesized CTPP tables as marginal totals. The perturbation of the CTPP results to address data disclosure issues may increase the likelihood of the "zero cell" problem, because the connection between the marginal and initial joint distribution will be further clouded.

For the second input, there is a research question of whether and how much to supplement the PUMS records with records from adjacent PUMAs. There is likely to be a tradeoff between data quality (or at least data consistency) and computing times.

Abbreviations used without definitions in TRB publications:

AAAE	American Association of Airport Executives
AASHO	American Association of State Highway Officials
AASHTO	American Association of State Highway and Transportation Officials
ACI–NA	Airports Council International–North America
ACRP	Airport Cooperative Research Program
ADA	Americans with Disabilities Act
APTA	American Public Transportation Association
ASCE	American Society of Civil Engineers
ASME	American Society of Mechanical Engineers
ASTM	American Society for Testing and Materials
ATA	American Trucking Associations
CTAA	Community Transportation Association of America
CTBSSP	Commercial Truck and Bus Safety Synthesis Program
DHS	Department of Homeland Security
DOE	Department of Energy
EPA	Environmental Protection Agency
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
FMCSA	Federal Motor Carrier Safety Administration
FRA	Federal Railroad Administration
FTA	Federal Transit Administration
HMCRP	Hazardous Materials Cooperative Research Program
IEEE	Institute of Electrical and Electronics Engineers
ISTEA	Intermodal Surface Transportation Efficiency Act of 1991
ITE	Institute of Transportation Engineers
NASA	National Aeronautics and Space Administration
NASAO	National Association of State Aviation Officials
NCFRP	National Cooperative Freight Research Program
NCHRP	National Cooperative Highway Research Program
NHTSA	National Highway Traffic Safety Administration
NTSB	National Transportation Safety Board
PHMSA	Pipeline and Hazardous Materials Safety Administration
RITA	Research and Innovative Technology Administration
SAE	Society of Automotive Engineers
SAFETEA-LU	Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (2005)
TCRP	Transit Cooperative Research Program
TEA-21	Transportation Equity Act for the 21st Century (1998)
TRB	Transportation Research Board
TSA	Transportation Security Administration
U.S.DOT	United States Department of Transportation

TRANSPORTATION RESEARCH BOARD
500 Fifth Street, N.W.
Washington, D.C. 20001

ADDRESS SERVICE REQUESTED

THE NATIONAL ACADEMIES™
Advisers to the Nation on Science, Engineering, and Medicine

The nation turns to the National Academies—National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council—for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org

ISBN: 978-0-309-22365-2