# Synthesizing Parcel-Level Households Using Census 2000 Data

Delores Muller, Southeast Michigan Council of Governments


## Abstract

Metropolitan planning organizations typically require individual household data for regional development forecasting and travel demand modeling. There is also an increasing demand for a full set of individual household data as input into disaggregated simulation models, such as UrbanSim and TRANSIMS. The costly expense of performing a detailed household travel survey, however, often results in a limited number of sampled households for small area analysis. To supplement Census 2000 data and household travel survey results, SEMCOG, the Southeast Michigan Council of Governments, has developed a low-cost procedure for synthesizing parcel-level household data to be used in small area forecasting. The procedure uses several Census 2000 data products and a series of Monte Carlo simulations to synthesize 11 characteristics for each household. The general approach is to prioritize the use of data sources by their coverage of household characteristics and synthesize characteristics that are consistent with aggregate block and block group level data while preserving multivariate distributions as represented by 5-Percent Public Use Microdata Sample data. Placement of the synthesized households into individual parcels is carried out using a ranked comparison of housing values and rental costs of the synthesized households to assessed property values in a digital parcel file. Households can then be analyzed at the parcel level or other levels of geography. SEMCOG has synthesized 125,000 households in Washtenaw County, Michigan. The synthesized households compare favorably with aggregate block and block group level data. Validation tests to determine whether the process results in reasonable multivariate distributions have yet to be completed.


## Introduction

As the metropolitan planning organization (MPO) for Southeast Michigan, SEMCOG produces a long-range regional development forecast (RDF) and travel demand forecast approximately every five years. Both forecasts require individual household data as input. Although a detailed household travel survey for the seven-county region has already been performed, the high expense of the survey results in a limited number of sampled households for forecasting at a smaller level of geography, such as traffic analysis zones (TAZs). Moreover, the forecast model UrbanSim will be used for producing SEMCOG's next RDF. SEMCOG's implementation of UrbanSim will require a full set of individual household data within 150 x 150 meter grid cells.

To supplement Census 2000 data and household travel survey results, SEMCOG has developed a low-cost procedure for synthesizing parcel-level household data to be used in small area forecasting. The procedure uses several Census 2000 data products and Monte Carlo simulations to synthesize individual household data with the following characteristics: household tenure, household type, sex of householder, age of householder, race of householder, presence of children, household size, household income, housing value or rental cost, number of vehicles available, and number of workers. The general approach of the procedure is to prioritize the use of data sources by their coverage of household characteristics and synthesize characteristics that are consistent with aggregate block and block group level data while preserving multivariate distr-ibutions as represented by 5-Percent Public Use Microdata Sample (PUMS) data.

As illustrated in the figure "Coverage of Census 2000 Data," information collected using the short form questionnaire is available as 100-percent data down to the block level in Summary File 1 (SF1). Similarly, long form sample data is available down to the block group level in Summary File 3 (SF3). Individual household data is available as 1- or 5-Percent PUMS. Although 1-Percent PUMS provides a fuller set of detailed household characteristics, 5-Percent PUMS provides individual household data at a smaller level of geography.

"The Synthesis Process" figure outlines the sequence of steps for synthesizing the household characteristics. In general, 100-percent data is synthesized at the block level before sample data are used to synthesize characteristics at the block group level. Several factors were considered when deciding the sequence and data to use for synthesizing individual characteristics. These factors include: coverage of the characteristic with respect to sample size and geographic detail, availability of multivariate data, degrees of freedom, and frequency of change in a characteristic.

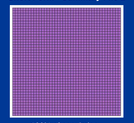
Given that cross-tabulated 100-percent data are available at the block level for the key life cycle household characteristics of household tenure, household type, sex of householder, and broad age of householder, the SF1 table H17 was chosen as the primary Census 2000 table to start with and add additional variables to. In situations where there is known homogeneity within a block, (e.g., the race of all householders in a particular broad age group is white), the known household information is assigned before synthesizing the remaining data. SF3 and 5-Percent PUMS data are used during the household synthesis process when additional multivariate distributions are required.

Once the synthesis of households has been completed, households may be placed into the individual parcels of a digital parcel file. With number of housing units assigned to each parcel, SF1 block level occupancy rates by tenure are applied to reduce the number of housing units that may be occupied. Because there are differences between Census 2000 and parcel data, there may not be an exact match in the number housing units and/or households at the block level. However, by controlling the number of housing units assigned at the parcel level to Census 2000 block group totals, the total number of households assigned to parcels can also be controlled to Census 2000 data at the block group level. In special circumstances where significant geocoding errors are apparent in Census 2000 data, block groups may be aggregated together for the purposes of controlling to Census 2000.

Placement of the synthesized households into individual parcels is carried out using a ranked comparison of housing values and rental costs of the synthesized households to assessed property values in a digital parcel file. Households can then be analyzed at the parcel level or other levels of geography by applying allocation and/or aggregation procedures

## Coverage of Census 2000 Data

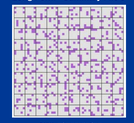
**Short Form - Summary File 1**



100% of population
Down to the block level

**Aggregate 100-Percent Data for:**
Tenure
Household Type
Age of Householder
Sex of Householder
Race of Householder
Presence of Children
Household Size

**Long Form - Summary File 3**



~16% of population
Down to the block group level

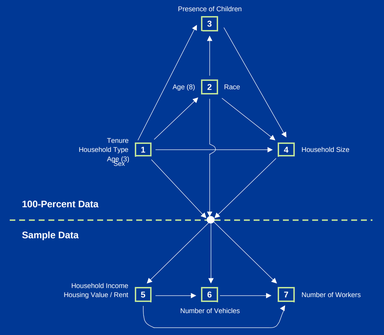
**Aggregate Sample Data for:**
Household Income
Housing Value / Rent
Number of Vehicles Available
Number of Workers
Other socio-economic data

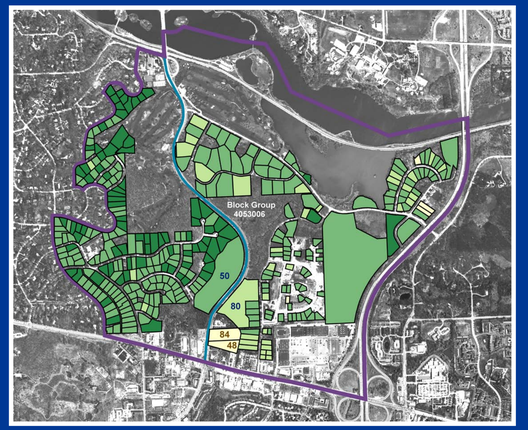
**5-Percent PUMS**

5% of population
In PUMAs (Areas with ~100,000 persons)

**Individual Household Data**
All data items

## The Synthesis Process




## Assessed property value (per housing unit)

- Under $100,000
- $100,000 to $249,999
- $250,000 to $499,999
- $500,000 or more
- #S Condominium units
- #S Apartment units
- TAZ boundary
- Block group boundary

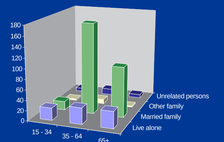
## Key Points

- Household synthesis should utilize the best data available.
- Placement of households into parcels is accomplished by comparing housing values and rental costs of synthesized households to assessed property values in a digital parcel file.
- Synthesizing parcel-level households is *not* an attempt to pinpoint and expose individual household data. The goal is to characterize neighborhoods using a set of realistic, yet synthetic household characteristics.

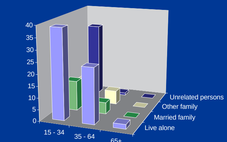
## Conclusions

Approximately 125,000 households in Washtenaw County, Michigan have been synthesized using this new procedure. The synthesized households compare favorably with aggregate block and block group level data. The figures below show that the synthesized households match the Census 2000 cross-tabulated data of household tenure, broad age of householder, and household type at the block group level. Although not explicitly shown, the synthesized households also match the cross-tabulated data at the block level. Validation tests to determine whether the process results in reasonable multivariate distributions have yet to be completed. Possible validation tests include chi-square and odds ratio tests of the synthesized households to a Census special tabulation or by obtaining access to Census 2000 1-in-6 individual long form data at a Census Research Data Center. Additional validation could be achieved by performing similar tests on households synthesized using iterative proportional fitting methods to compare test results.

Owner Households by Age of Householder and Household Type




Renter Households by Age of Householder and Household Type




SEMCOG plans to synthesize an additional 1.7 million households using this procedure. The open source code will likely be ported from Visual Basic to Python for faster execution during the next year.