San Francisco Travel Demand Forecasting Model Development

Population Synthesis

Final Report

prepared for
San Francisco County Transportation Authority

prepared by
Cambridge Systematics, Inc.

Updated by:
San Francisco County Transportation Authority

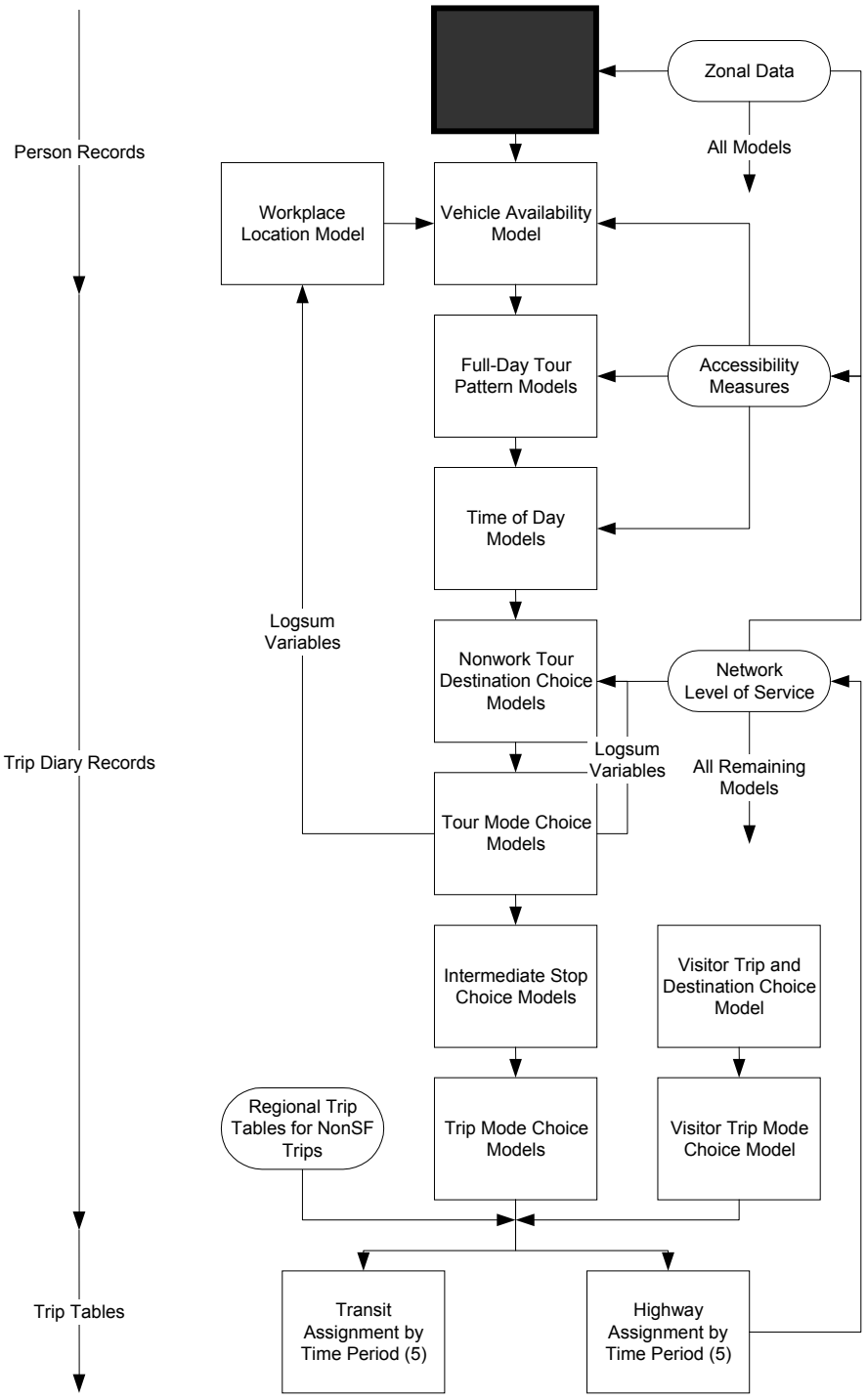
October 1, 2002

Table of Contents

Introduction

This report describes the methods and software for generating a prototypical sample of households for San Francisco County for the base year 1998 and various forecast years. Figure 1 shows where Population Synthesis fits in the Model Structure.

Figure 1. San Francisco Model System

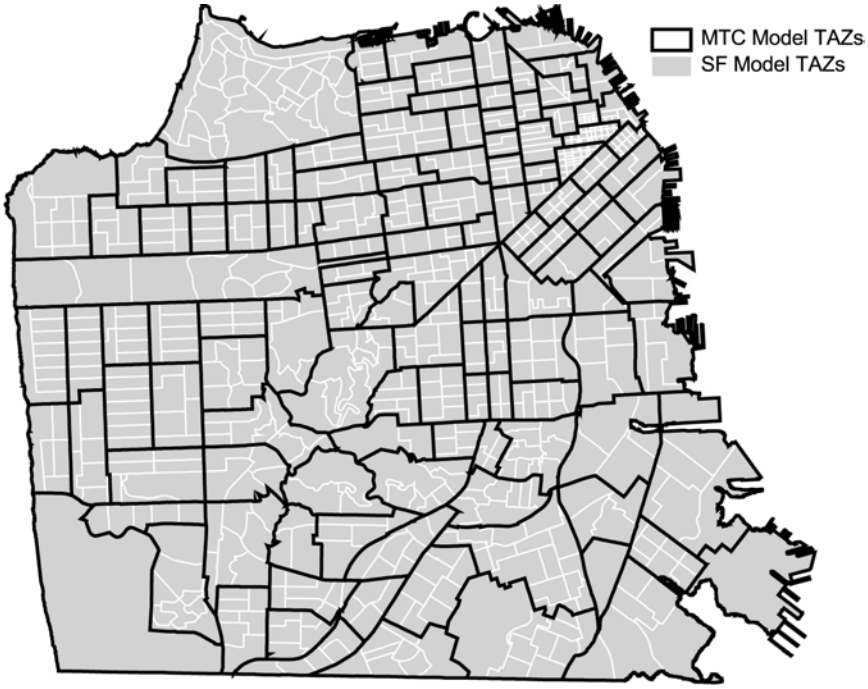


Regions

For San Francisco County, the procedure uses a hierarchy of three region/zone systems:

- 6 PUMAs (Public Use Microdata Area), containing
- 127 MTC TAZs (Metropolitan Transportation Commission Transportation Analysis Zones (MTAZs)), containing
- 766 SFCTA TAZs (San Francisco County Transportation Authority Transportation Analysis Zones (SFTAZs)).

Figure 2. Map of San Francisco and MTC Travel Analysis Zones (TAZs)



Sampling Cell Categories

To classify sampling "cells", marginal distributions along the following dimensions are used:

- A: Household size / number of workers
  - A1: 1 person / 0 workers
  - A2: 1 person / 1 worker
  - A3: 2 persons / 0 workers
  - A4: 2 persons / 1 worker
  - A5: 2 persons / 2 workers
  - A6: 3+ persons / 0 workers
  - A7: 3+ persons / 1 worker
  - A8: 3+ persons / 2 workers
  - A9: 3+ persons / 3+ workers
- B. Household income (in 1990 dolloars)
  - B1: Less than $20,000
  - B2: $20,000 - $39,999
  - B3: $40,000 - $74,999
  - B4: $75,000 or more
- C. Age of head of household
  - C1: Under 35
  - C2: 35 – 61
  - C3: 62 or older

Across the three marginals combined, there are 9 x 4 x 3 = 108 different sampling cells.

The nine categories for household size/number of workers were chosen because they efficiently distinguish between important household lifecycle groups. The specific breakdowns for income and age were chosen because they correspond to categories that are available in the MTC future year land use files, so updating the populations to future years can be kept consistent with MTC breakdowns within zones. Also, all of these categorizations are compatible with the Census tables available in the Census Transportation Planning Package (CTPP) Urban Element.

Sampling Process

Census Year

Step 1- Classifying PUMS (Public Use Microdata Sample) households

The first step in sampling is to take the 1990 PUMS 5% Microdata sample, and to classify each household according to PUMA and the marginal categories described above. This program produces the following outputs:

- A: A count of the number of PUMS households for each PUMA/cell (6 x 108).
- B: A file of PUMS household ID numbers, sorted by PUMA and cell.
- C: The PUMS records rewritten in a binary file format, to allow fast file access and manipulation in subsequent sampling programs.
- D: Counts of the numbers of households, persons and workers in each marginal category in each PUMA. These numbers are used to derive some details that are not available in the Census tables – particularly:
  - The percentage of people in each age category who are head of the household;
  - The average number of people in the 3+ person household type categories;
  - The average number of workers in the 3+ worker household type category.

Step 2 – Calculating marginal distributions from the Census

The census data available in CTPP includes tables for each of the marginal distributions above for each Census Travel Analysis Zone (CTAZ), in tables U111, U112, U113 and U114. The program to process these tables performs the following functions:

- Collapse the Census table categories for household size, number of workers, income and age into our categorization described above.
- Add group quarter households (U112) together with family households (U111).
- Adjust the population totals by age to include only heads of households; using the PUMA-specific fractions calculated from PUMS data in Step 1.
- Produce marginal tables for 1990 at the PUMA, MTAZ and SFTAZ level.

Base and Forecast Years

Step 1 –Deriving marginals at the SFTAZ level

For a base year or forecast year, the program expects the user to provide a population forecast for each SFTAZ for each of the following three items:

- the number of households residing in the SFTAZ
- the number of people residing in the SFTAZ
- the number of employed people residing in the SFTAZ

The sampling program needs to use this input, along with the best available additional data, to determine the marginals along household type, income and age of head of household for each SFTAZ. This is done as follows:

ABAG (Association of Bay Area Governments) produces population forecasts at the Census tract level. MTC has already used those forecasts to produce base year and forecast year land use files at the MTAZ level, which is roughly equivalent to the Census tract level within SF County. For efficiency's sake, and to maintain consistency with MTC, the MTC files are used to calculate fractional changes between the base year and the forecast year for each marginal category for each MTAZ.

For household income, our categorization is the same as that in the MTC files, so the calculation is very simple – for each MTAZ, just divide the forecast year households in each income category by the base year households to derive the growth factor for that category.

For age of head of household, the MTC files contain forecasts of the fraction of the population aged 62 or over. With this information, we can calculate the growth factors for both the population with age under 62 and the population with age 62+. If we make the assumption that the growth factor is the same for the populations under age 35 and age 35-61, then this is sufficient. To do better than that, we would need to go back to the ABAG projections. That was judged not to be worth the extra time and effort, since most behavioral differences between these two age groups are due to the number of children and workers in the household, and the amount of income available, and we are controlling for those factors. Purely age-related behavioral differences, such as difficulties walking and driving, become important with the older age group.

For household size and number of workers, the MTC files (and ABAG data) only give the total numbers of residents and employed residents within each MTAZ, but do not provide any distributions. Since these are the same variables we have at the SFTAZ level, we do not use any MTAZ-level information for this step. Instead, to

estimate an appropriate household type distribution for a given SFTAZ, the following iterative procedure is used:

- **Substep 1-A:** The census year or base year marginal fractions from 1990 are used as starting values, and the total number of persons is calculated. (This requires using the PUMS-derived average household sizes for the categories with 3+ persons in the household).
- **Substep 1-B:** If the number of persons in the SFTAZ is lower than the forecast year population, then a fraction of the 1-person households are moved into the 2 and 3+ person household categories, proportional to existing fractions. Otherwise, if the number of persons is too high, the fractions are shifted in the opposite directions.
- **Substep 1-C:** The new marginal fractions are used to calculate the number of people in the SFTAZ (as in Step A). Steps B and C are iterated until the total population matches the SFCTA forecast year projections for the zone.
- **Substep 1-D:** Using the marginal fractions from steps A to C as initial values, calculate the number of workers in the SFTAZ. (This requires using the PUMS-derived average workers for the 3+ worker category).
- **Substep 1-E:** If the number of employed residents in the zone is lower than the forecast year projections, then a fraction of the 1 person/0 worker households are shifted to 1 person/1 worker households, and a fraction of the 2 and 3+ person/ 0 and 1 worker households are shifted to the corresponding multiple-worker categories, proportional to existing fractions. Otherwise, if the number of workers is too high, then the fractions are shifted in the opposite directions.
- **Substep 1-F:** The new marginal fractions are used to calculate the number of employed residents in the SFTAZ (as in Step D). Steps E and F are iterated until the employed population matches the SFCTA forecast year projections for the zone.

The whole procedure (steps B to F) is iterated three times, enough to ensure that both the number of total residents and employed residents are matched simultaneously as closely as possible for the zone.

Step 2 - Iterative proportional fitting of cells

Instead of just estimating the marginal distributions within each SFTAZ (9+4+3=16 values), we want to estimate the number of households within each cell in each SFTAZ (9*4*3=108 values). This is typically done with iterative proportional fitting (IPF), iterating on the three different marginal distributions and adjusting the cells in each until all three distributions are matched simultaneously. For the base year, we also apply the 2-stage approach proposed by Beckman, et al. ("Creating Synthetic Baseline Populations", Paper LA-UR 95-1985). The first stage uses IPF of the cell distribution of the PUMS sample households within a PUMA against the marginals summed across all SFTAZs within that PUMA. This generates a

"maximum information" target for each cell in the PUMA, which is used as an additional marginal target in the second stage SFTAZ-level IPF.

For the forecast year, we begin with the base year results as starting values (replacing any 0's with 1's), and then use IPF to match the forecast year marginals for each zone.

This process runs quickly and reaches equilibrium after about 30 iterations for both base year and forecast year distributions. Before using here, the program was tested on the example data given in the Beckman, et al. paper.

Step 3 - Drawing PUMS households to match the IPF results

Once we know the appropriate number of households from each of the 108 cells for a given SFTAZ, we go to the PUMS households for the appropriate PUMA and randomly sample the correct number of households within each cell. This program draws a PUMS household ID for each household in the prototypical sample. Because the PUMS is a 5% sample, each PUMS household will appear in the full sample about 20 times in the base year. The same PUMS household could be sampled more than once within a given SFTAZ, but will be generally be sampled for a number of different SFTAZs within the PUMA. In future years, as the total number of household grows, each PUMS household will appear more than 20 times on average, but some may be included less often if those particular types of households are predicted to become less common in the population.

There are a couple of complications in the program to draw the PUMS households:

- The IPF produces fractional numbers of households in cells, while we wish to draw integer numbers of households. For example, if the IPF indicates we should sample 2.45 households in a given SFTAZ/cell, then we should probably sample either 2 or else 3. This problem is solved using a stochastic procedure based on the fractional remainder, to get the correct expected number of households. For the example of 2.45 households, there is a 0.45 probability that we sample 3, or a 0.55 probability that we sample 2. (The expected value is thus 2 * .55 + 3 * .45 = 2.45.) This procedure is done for each cell within an SFTAZ, and then a check is made that the total number of households drawn across all cells is correct. If not, then the process is iterated until the total is correct.
- Ideally, the variation between two samples should be a function of the different socio-demographic forecasts, and should be influenced as little as possible by random sampling error. In order to control this, a file is generated which contains a different random number generator seed for each SFTAZ/cell combination. (These seeds themselves are generated as a random sequence.) When drawing PUMS households, the program returns to the same random number seed each time it draws for a given SFTAZ/cell. This means that it will generate the same sequence of households each time for that SFTAZ/cell, only the number drawn will vary. Thus, if there are 60 households in cell 7 for

SFTAZ 614 in 1998, and 70 households for that cell/SFTAZ in 2020, then the first 60 households will be identical for the two years, but a new 10 will be added on for 2020.

The prototypical sample file that is generated contains the following data items for each sampled household:

- The SFTAZ number
- The PUMA number
- The sampling cell number (1 to 108)
- The PUMS household ID number.

Step 4 – Attaching the PUMS records

To use the samples in application, we need to go back to the full PUMS data file and use the household ID numbers to append the relevant household and person variables that are used in the travel demand models. This is done by the final step in the sampling program. The final output format is described in the Appendix.

Sampling Results

Table 1 and Figures 3 through 6 show the input values totaled across all zones, and the resulting marginals for each forecast year. The numbers of households, residents and employed residents are summed from the SFCTA zonal input files. The marginals are also summed across zones after the marginals are calculated for each year.

To match a pattern of dropping household size and increasing workers per household, the 1-person/1 worker and 2 person/2 worker household types are the only ones that grow throughout the period.

The lowest income group shrinks across the period, while the highest income groups grow substantially. In about 2005, the size of the 4 groups becomes about equal, and by 2020, the highest income group is the largest.

Up until about 2010, most of the growth is in households with the head aged 35-61. After 2010, the number of older households (head aged 62+) begins to grow sharply, and the number of younger households begins to drop. (Remember that the age and income trends come from ABAG projections at the Census block level.)

Table 1. Summary of Sample Population by Year

Year	1998	2000	2005	2010	2015	2020
Households	313,991	317,572	323,854	330,619	333,959	336,290
Residents	742,012	756,700	764,931	775,261	771,255	761,981
Employed	390,860	400,962	426,410	457,253	462,559	471,969
Residents/HH	2.36	2.38	2.36	2.34	2.31	2.27
Workers/HH	1.24	1.26	1.32	1.38	1.39	1.40
Households by Persons/Workers						
1 P / 0 W	47,058	45,887	44,444	42,573	43,342	43,639
1 P / 1 W	72,293	73,541	80,361	88,131	93,045	99,368
2 P / 0 W	20,484	20,255	18,815	16,855	15,881	14,284
2 P / 1 W	23,621	23,550	22,095	19,984	18,978	17,142
2 P / 2 W	44,629	45,380	50,545	56,530	59,665	63,241
3+P / 0 W	13,879	14,356	12,954	11,787	11,091	10,183
3+P / 1 W	26,849	27,434	25,065	22,320	20,802	18,505
3+P / 2 W	33,800	34,319	32,107	29,190	27,622	24,881
3+P / 3+W	31,376	32,850	37,467	43,249	43,532	45,049
Households by Income Group						
Inc <25 K	98,256	93,980	82,218	67,122	61,303	58,046
Inc 25-45	79,009	78,243	76,272	76,429	78,231	78,303
Inc 45-75	74,266	77,043	83,563	90,293	93,520	95,723
Inc >75K	62,448	67,633	79,783	92,114	96,195	99,212
Households by Age of Head						
Age < 35	84,261	85,028	88,415	90,719	89,150	82,473
Age 35-61	140,875	141,895	146,952	150,536	147,431	135,250
Age 62+	88,854	90,649	88,488	89,364	97,378	118,569

Figure 3. Sample Population by Year

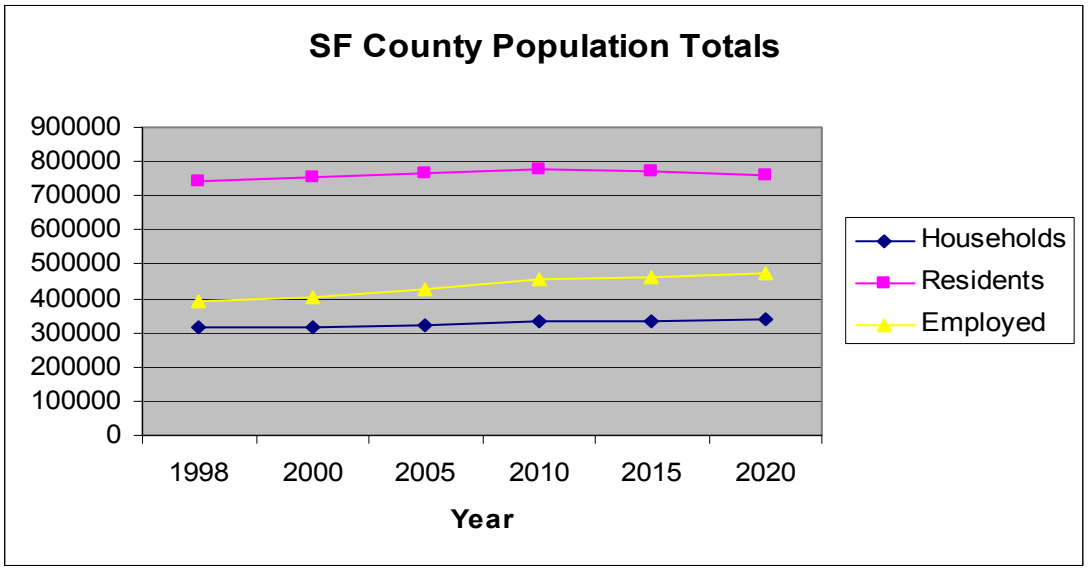


Figure 4. Sample Households by Size and Worker and Year

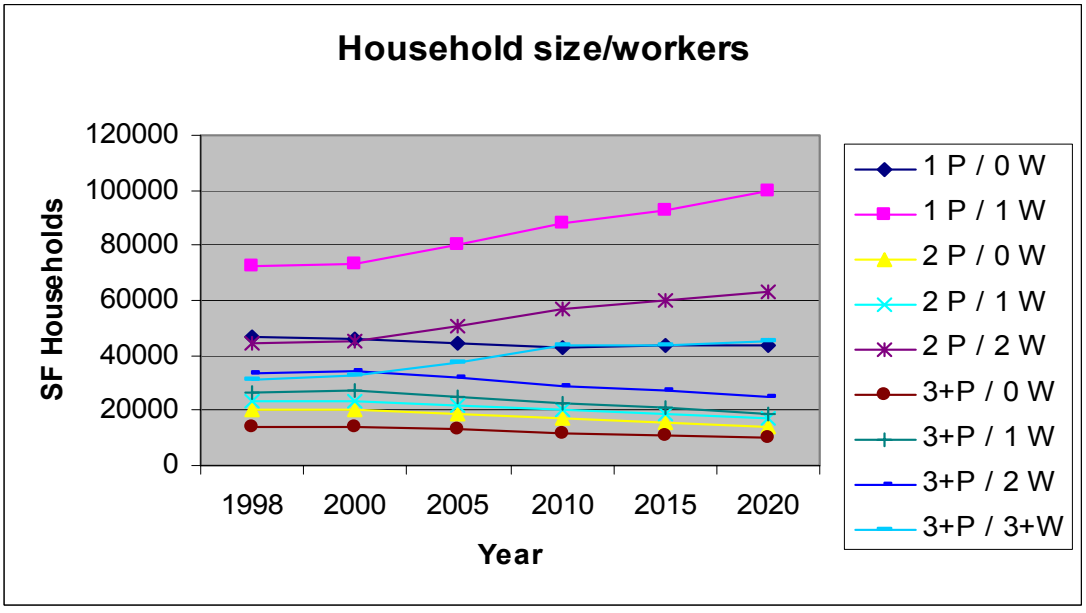


Figure 5. Sample Households by Income and Year

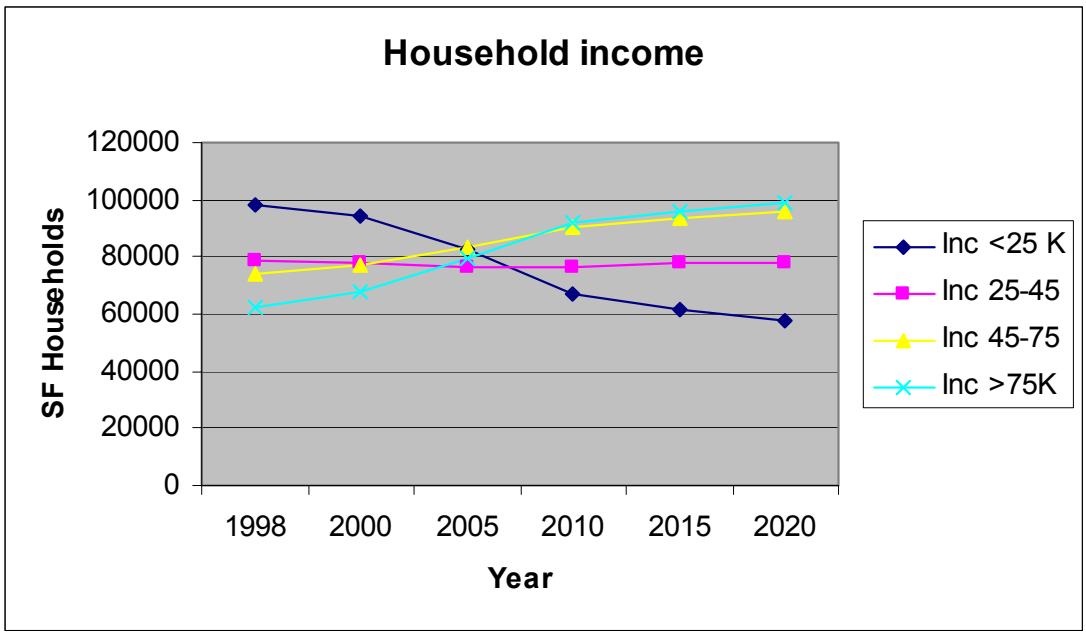


Figure 6. Sample Households by Age of Head of Household and Year