WEIGHTING THE 1996 AND 1997 AMERICAN COMMUNITY SURVEYS[1]

Scot A. Dahl, U.S. Bureau of the Census, Washington DC 20233

Keywords: American Community Survey; weighting

1.0 Overview of the American Community Survey

The American Community Survey is an annual survey, under development to provide demographic information about communities and populations every year. This survey will collect the data traditionally collected by the decennial census long form. Development started in 1996 in four sites: Rockland County, NY; Brevard County, FL; Multnomah County, OR; and Fulton County, PA. In 1997, four more sites were added: Douglas County, NE; Otero County, NM; Franklin County, OH, and Houston, TX. Though it is an annual survey, the questionnaires are mailed out, and the data collected, monthly.

Eventually, in 2003, the survey will be in full-scale production, mailing 250,000 questionnaires per month to every county in the nation. Annual profiles will be produced for all states, cities, counties, metropolitan areas, or population groups for 65,000 or more people. For smaller areas, two to five years of data will be accumulated to produce estimates similar to those of the census long form.

1.1 Sampling Procedure

For each site, a systematic sample was drawn from the Master Address File (MAF) for the site. The MAF is a file of all addresses in a county developed from Census Bureau and US Post Office address listings. The 12-month sampling rates for both small governmental units (SGUs), those with less than 2,500 population, and non-SGUs are shown in table 1A. One twelfth of the selected sample was mailed each month.

Table 1A: ACS Base Sampling Rates: 1996 & 1997

Site	1996		1997	
	Non-SGU	SGUs	Non-SGU	SGUs
Brevard, FL	0.15	0.30	0.03	0.09
Rockland, NY	0.15	0.30	0.03	0.09
Multnomah,	0.15	0.30	0.03	0.09
Fulton, PA	na	0.30	na	0.09
Douglas, NE			0.15	0.30
Otero, NM			0.03	0.09
Franklin, OH			0.03	0.09
Houston, TX			0.03	0.09

1.2 Data Collection

Three data collection modes were used to conduct the 1996 and 1997 American Community Surveys: Mail, Computer Assisted Telephone Interviewing (CATI), and Computer Assisted Personal Interviewing (CAPI). These three modes are described below.

Mail Phase: The Mail phase began with a pre-notice letter mailed to each sample address on the second to last Wednesday of the month preceding the sample month. The ACS Questionnaire was mailed one week later. One week after that, a reminder card was mailed to all sample housing units. A replacement questionnaire was mailed two weeks later if the original questionnaire had not yet been checked in.

CATI Phase: Approximately five weeks after the mailing of the initial ACS questionnaire, the CATI staff began contacting nonresponding sample addresses by telephone. This phase lasted for approximately one month.

CAPI Phase: The CAPI universe consisted of all outstanding non-response cases remaining after the completion of the CATI phase. A 1 in 3 subsample was selected from these outstanding cases and forwarded to the Field Representatives. Field Representatives visited each assigned household and conducted an interview. The CAPI phase also lasted for approximately one month. For Otero County, NM some addresses were un-mailable. Two-thirds of these cases were selected and sent directly to CAPI.

As an illustration, for Rockland County NY in February 1997, 251 questionnaires were mailed for panel 9702. Of these, 114 returned as mail responses in February 1997, 56 were CATI or late mail returns received in March, and 81 were late mail returns in April or were subjected to CAPI subsampling.

2.0 Description of the Source Files

Sample Address/Control File: The address/control file contained the status and outcome codes for every housing unit address included in the ACS sample. This file also contained the geographic codes (tract, block, address, etc.) and sampling stratum for every sample address.

Master Address File (MAF): This was the sampling frame. It contained geographic and other information for every address in each site. This file was constructed from the 1990 Census Address Control File (ACF), modified by periodic deliveries of the USPS Delivery Sequence File (DSF). The 1996 and 1997 samples were originally

drawn in September of 1995 and 1996 respectively. After data collection for a sample year was completed, a later version of the MAF was used:

- To update the sample's original geographic codes
- To produce tract level housing unit counts to be used in the final stages of the weighting.

Edited Data Files: The edited data files contained the edited response data. The edits handled item non-response by imputing from other reported information or by hot-deck imputation. For each state the edited data files consisted of two subsets: one file of responses to housing unit questions, the other of population responses.

Population Control Counts: This file contained the most recent population estimates for the counties in the ACS survey. These independent estimates were produced using demographic analysis by the Census Bureau's Population Division and consisted of housing unit (non-Group Quarters) population estimates broken down by:

- Age (one year intervals to age 85)
- Sex (Male, Female)
- Race (White, Black, American Indian, API)
- Ethnicity (Hispanic/Non-Hispanic)

3.0 Preliminary Operations

3.1 MAF Operations

As mentioned above, two operations were performed with the most current version of the MAF available after data collection was completed (around February of 1997 and 1998 respectively for the 1996 and 1997 surveys). These operations were to:

- Update the geography of the original sample.
- Produce counts of valid housing units by tract for later use in the final stages of the weighting.

3.2 Edits and Record Selection

The edits were applied to the raw housing unit and population response data. While not strictly a part of the weighting process, certain edits had to be completed in order to weight each record. The edits imputed for item nonresponse using other data reported by the household, or if necessary, by substituting a value from a nearby neighbor (hot deck imputation).

In the event of multiple responses from one sampled address, the Record Selection Algorithm selected the response to be retained. It also reclassified some records as nonresponses and, rarely, created additional person records for a household. As part of the record selection process: a status (occupied, vacant, delete, etc.), a mode (Mail, CATI, or CAPI), and a tabulation month assigned to every housing unit address, both responses and nonresponses.

3.3 Creation of the Initial Weighting Files

Two initial weighting files were created for each site: a housing unit file with one record for each sampled housing unit address, and a population file with one record for each person in responding housing units.

3.4 Disclosure Avoidance Data Swapping

Some housing units and their members, with characteristics unique within their block group, were swapped with similar housing units in other block groups within the same site. This was done to reduce the possibility that any information about an individual housing unit or its members could be deduced from the tables and the public use files produced. Approximately 1% of the occupied housing units in each site were swapped.

4.0 Initial Housing Unit Weighting Factors

The 1996 and 1997 surveys were processed separately but basically the same set of weighting factors was used. The weighting factors used in the 1996 and 1997 ACS fell into four general categories:

- Base weight and mode adjustments (BW, SSF, VMS)
- Sample size factors (SRF, FAF)
- Nonresponse factors (NIF1, NIF2, MBF)
- Post-stratification factors (HPF1, HPF2, PPSF, PPF)

4.1 Base Weight (BW)

This weight was assigned to every housing unit address and was basically the inverse of an address's sampling rate. An adjustment was made to account for housing units on the MAF that had already been sampled by other Census Bureau surveys, including the 1996 ACS.

The table below shows the ranges of base weights assigned for Small Governmental Unit (SGU) and non-SGU areas of each site for 1997. The base weights in 1996 were 3.33 for SGUs and 6.67 for non-SGUs. The wide range of values in 1997 for the four 1996 sites reflects the adjustment made to unduplicate the 1997 sample from the 1996 sample.

Table 4A: Range of Base Weights: 1997

Site	1997	
	Non-SGU	SGUs
Brevard, FL	27.4 - 38.5	7.44 - 11.5
Rockland, NY	27.2 - 42.4	7.39 - 11.5
Multnomah, OR	27.5 - 40.3	7.16 - 11.2
Fulton, PA	Na	7.27 - 11.2
Douglas, NE	6.63 - 6.67	3.33 - 3.35
Otero, NM	33.1 - 33.3	11.1
Franklin, OH	33.2 - 33.3	11.1 - 11.2
Houston, TX	33.2 - 33.3	11.1 - 11.2

4.2 Sample Reduction Factor (SRF)

This factor was used for the 1997 sample only and affected only the addresses in the four 1996 sites that were mailed in 1996 and tabulated in 1997. In the four 1996 sites, the sample sizes were reduced from approximately 15% and 30% to 3% and 9% in 1997. Since the survey is based on the month a response was received (not when the questionnaire was mailed), January and February 1997 would have had unusually large numbers of responses compared to the other ten months of 1997. To reduce the possible bias to the editing and weighting, these "carried-over" records were sub-sampled down to the standard 1997 rates. The sample reduction factor reflects the increased weight assigned to records that were retained.

4.3 CAPI Subsampling Factor (SSF)

This factor was 1.0 for all Mail and CATI cases, 3.0 for those records selected in CAPI subsampling, and zero for those not selected. The actual assignment of a value for SSF was somewhat complicated by late mail returns received during the month of CAPI operations. For Otero County NM, some addresses were unmailable. A two-thirds sample of these were sent directly to CAPI and for these cases SSF = 1.5.

4.4 Variation in Monthly Response by Mode (VMS)

This factor made the total weight of the Mail, CATI, and CAPI subsampled records tabulated in a month equal to the total weight of all cases originally mailed for that month. Twelve factors were computed for each site, one for each month of the year. VMS for site *s* in month *i* was defined as:

$$VMS_{si} = \frac{M_{si} - MAIL_{si}}{CATI_{si} + CAPI_{si}}$$

Where for site *s*:
M_{si} = the weight of all questionnaires mailed in month *i*
$MAIL_{si}$ = the weight of mail responses in month *i*
$CATI_{si}$ = the weight of CATI responses in month *i*
$CAPI_{si}$ = the weight of the CAPI responses in month *i*
s = site code (1, 2, ... , 8)
i = month code (1, 2, ..., 12)

This value of VMS was applied to all CATI and CAPI housing units in site *s*, in tabulation month *i*. For all Mail responses, VMS = 1.0. For addresses whose questionnaires were mailed in November or December but not received until the following year, VMS = 0. The table below shows the range of values of VMS for the eight sites.

Table 4B: Range of Values of VMS: 1996 & 1997

Site	1996	1997
Brevard, FL	0.89 - 1.13	0.82 - 1.29
Rockland, NY	0.91 - 1.13	0.72 - 1.43
Multnomah, OR	0.78 - 1.16	0.86 - 1.26
Fulton, PA	0.79 - 1.28	0.77 - 1.87
Douglas, NE		0.88 - 1.19
Otero, NM		0.74 - 1.36
Franklin, OH		0.86 - 1.29
Houston, TX		0.91 - 1.14

4.5 Noninterview Factor #1 (NIF1)

This was the first of two factors that adjusted the weight of all respondents to account for both respondents and nonrespondents. NIF1 was a tract-based nonresponse adjustment; it increased the weight on responding housing units to account for both respondents and similar nonrespondents. "Similar" here meant within the same tract and of the same building type (single or multi-unit).

$$NIF\,1_{cbt} = \frac{RES\,1_{cbt} + NON\,1_{cbt}}{RES\,1_{cbt}}$$

where:
RES1 = weighted sum of all responding housing units
NON1 = weighted sum of the non-respondents
c = county code
b = building type (single or multi-unit)
t = tract code

This value of NIF1 was applied to all occupied, respondent housing units in county *c*, tract *t*, of building type *b*. For vacancies and non-housing units (businesses, vacant lots, etc.), NIF1 = 1.0. For nonrespondents, NIF1 = 0. The range of values of NIF1 for the eight sites is shown below.

Table 4C: Range of Values of NIF1: 1996 & 1997

Site	1996	1997
Brevard, FL	1.00 - 1.16	1.00 - 1.32
Rockland, NY	1.00 - 1.11	1.00 - 1.36
Multnomah, OR	1.00 - 1.22	1.00 - 1.24
Fulton, PA	1.00 - 1.05	1.00 - 1.02
Douglas, NE		1.00 - 1.29
Otero, NM		1.00 - 1.06
Franklin, OH		1.00 - 1.24
Houston, TX		1.00 - 1.69

4.6 Non-Interview Factor #2 (NIF2)

The second nonresponse adjustment factor was NIF2. This factor was computed in the same manner as NIF1 except that similar housing units were now defined as being tabulated in the same month instead of in the same tract, and NIF2 was computed given that NIF1 has already been applied.

$$NIF\ 2_{sbm} = \frac{RES\ 1_{sbm} + NON\ 1_{sbm}}{RES\ 2_{sbm}}$$

Where:
RES1 = weighted sum of all responding housing units
RES2 = weighted sum of all responding housing units (computed using NIF1)
NON1 = weighted sum of the non-respondents
s = site code
b = building type (single or multi-unit)
m = tabulation month

This value of NIF2 was applied to all responding occupied housing units in site *s*, tabulation month *m*, of building type *b*. For vacancies and non-housing units, NIF2 = 1.0. For nonrespondents, NIF2 = 0. The range of values of NIF2 for the eight sites is shown below.

Table 4D: Range of Values of NIF2: 1996 & 1997

Site	1996	1997
Brevard, FL	0.98 - 1.02	0.99 - 1.04
Rockland, NY	0.98 - 1.03	0.92 - 1.09
Multnomah, OR	0.98 - 1.04	0.97 - 1.03
Fulton, PA	0.98 - 1.04	0.99 - 1.10
Douglas, NE		0.98 - 1.04
Otero, NM		0.97 - 1.17
Franklin, OH		0.98 - 1.05
Houston, TX		0.97 - 1.05

4.7 Mode Bias Factor (MBF)

This factor was an attempt to compensate for the bias resulting from not taking the mode of response into account when calculating NIF1 and NIF2. The concern was that there were systematic differences between the households that responded by mail and those that did not.

The first step in computing MBF was to calculate an alternative noninterview adjustment, NIFM, using only the CAPI respondents in the denominator. The underlying assumption was that the characteristics of nonrespondents were most like those of the hardest-to-get respondents.

4.7a Noninterview Factor - Mode (NIFM)

This factor was similar to NIF2 in that housing units were grouped by tabulation month (as well as by building type). However NIFM adjusts the weight of just the CAPI respondents to account for both CAPI respondents and all nonrespondents. MAIL and CATI cases receive a value of NIFM = 1.0. This factor was not applied directly but rather as part of computing MBF. NIFM was computed:

$$NIFM_{sbm} = \frac{RESM_{sbm} + NONM_{sbm}}{RESM_{sbm}}$$

where:
RESM = weighted sum of the CAPI respondents
NONM = weighted sum of all non-respondents
s = site code
b = building type (single or multi-unit)
m = tabulation month

For purposes of computing MBF, this value of NIFM was applied to all CAPI responding occupied housing units in site *s*, tabulation month *m*, of building type *b*. For Mail and CATI respondents, and for vacancies and non-housing units, NIFM = 1.0. For nonrespondents, NIFM = 0. The range of values of NIFM for the eight sites is shown below.

Table 4E: Range of Values of NIFM: 1996 & 1997

Site	1996	1997
Brevard, FL	1.00 - 1.19	1.00 - 1.28
Rockland, NY	1.00 - 1.21	1.00 - 1.29
Multnomah, OR	1.00 - 1.27	1.00 - 1.20
Fulton, PA	1.00 - 1.18	1.00 - 1.08
Douglas, NE		1.00 - 1.27
Otero, NM		1.00 - 1.07
Franklin, OH		1.00 - 1.14
Houston, TX		1.00 - 1.19

4.7b Computing MBF

This factor made the total weight, within specified weighting cells, the same as if NIFM had been used instead of NIF1 and NIF2. For any specified group of housing units, the total weight could now be computed two ways. One was to use the nonresponse adjustments NIF1 and NIF2. The other way was to use the other nonresponse adjustment NIFM. MBF was the factor that, when applied to NIF1 and NIF2, caused these two results to be equal for a specific grouping of housing units. For computing MBF, housing units within a site were grouped by tenure, tabulation month, and the marital status of the householder. For each group two weighted totals were computed, one using NIF1 and NIF2, the other using NIFM. MBF is computed as the ratio of these.

$$MBF_{somr} = \frac{M_{somr}}{N_{somr}}$$

where:
N = weighted sum computed using NIF1 x NIF2
M = weighted sum for the same cell computed using NIFM
s = site code
o = tenure (owner or renter)
m = tabulation month
r = marital status

This value of MBF was applied to all responding occupied housing units in site *s*, with tenure *o*, in tabulation month *m*, with marital status *r*. For vacancies and non-housing units, MBF = 1.0. After MBF was applied, the weight of a housing unit was:

HU Wgt = BW x SRF x SSF x VMS x NIF1 x NIF2 x MBF

Note that NIFM was not used directly; it was included indirectly as part of the computation of MBF.

The range of values of MBF for the eight sites is shown below. In general, MBF was higher for temporarily occupied housing units than for permanently occupied ones. It also tended to be higher for renters than for owners.

Table 4F: Range of Values of MBF: 1996 & 1997

Site	1996	1997
Brevard, FL	0.99 - 1.05	0.97 - 1.07
Rockland, NY	0.99 - 1.03	0.94 - 1.07
Multnomah, OR	0.98 - 1.03	0.98 - 1.03
Fulton, PA	0.98 - 1.01	0.92 - 1.04
Douglas, NE		0.99 - 1.05
Otero, NM		0.98 - 1.01
Franklin, OH		0.99 - 1.01
Houston, TX		0.99 - 1.04

4.8 Furlough Adjustment Factor (FAF)

This factor only applied to the 1996 survey. It adjusted the weights of the February 1996 CAPI records to account for the "missing" January 1996 CAPI cases caused by the furlough of late 1995/early 1996. This value was approximately 2.0 for February CAPI cases and exactly 1.0 for all others.

4.9 Housing Post Stratification Factor #1 (HPF1)

By the time data collection for a sample year was completed, the original sample was almost a year-and-a-half old. In order to account for any new construction that had occurred since the sample was selected and to reflect any geographic coding changes, the weighted number of housing units was compared to housing unit counts from a newer Master Address File (MAF). The factor HPF1 made the weighted number of housing units in a tract equal to the MAF housing unit count for the tract. HPF1 is defined as:

$$HPF\,1_{ct} = \frac{MAFHU_{ct}}{HU_{ct}}$$

where:
MAFHU = a count of MAF addresses in county *c*, tract *t*
HU = weighted estimate of housing units in county *c*, tract *t* prior to applying HPF1
c = county code
t = tract code

This value of HPF1 was applied to all housing unit addresses (including vacancies and non-housing units) in county *c*, tract *t*. The range of values of HPF1 for the eight sites is shown below.

Table 4G: Range of Values of HPF1: 1996 & 1997

Site	1996	1997
Brevard, FL	0.92 - 2.45	0.71 - 1.47
Rockland, NY	0.96 - 1.36	0.82 - 1.35
Multnomah, OR	0.60 - 1.94	0.83 - 1.80
Fulton, PA	0.98 - 1.02	0.96 - 1.04
Douglas, NE		0.95 - 1.33
Otero, NM		0.94 - 1.17
Franklin, OH		0.66 - 1.74
Houston, TX		0.62 - 2.29

5.0 Person Post Stratification Factor (PPSF)

After computing HPF1 for housing units, the weights of the persons in the housing units were computed. Initially, each person in a housing unit was assigned the weighting factors (BW, ..., HPF1) of their associated housing unit. Then an iterative process was run to compute PPSF.

This factor was then applied to individual person records based on their age, race, sex and Hispanic origin. It adjusted person weights so that the weighted sample of persons approximately matched county population control counts by age, race, sex, and Hispanic origin. These control counts were provided by the Census Bureau's Population Division and reflect an estimate of the population of the county at July 1st.

The first iteration adjusted for race, sex and age group.

$$PPSF(R)_{rsa} = \frac{CC_{rsa}}{POP_{rsa}}$$

Where:
CC = Control counts for race *r*, sex *s*, age group *a*
POP = ACS estimate for race *r*, sex *s*, age group *a*
r = race code (White, Black, Other)
s = sex code (Male, Female)
a = age group (a grouping of five or more ages, such as: 0-4, 5-9, 10-19, etc.)

After PPSF(R) was applied to all person weights, a second iteration was run that adjusted the weighted population to match the control counts by Hispanic origin, sex and age group. (The age groups used for the Hispanic origin adjustment may have been different than the ones used for the race adjustment).

$$PPSF(H)_{hsa} = \frac{CC_{hsa}}{POP_{hsa}}$$

Where:
h = Hispanic code (Hispanic, Non-Hispanic)

PPSF(R) was then recomputed after all previously computed values of PPSF(R) and PPSF(H) from the prior iterations had been applied to each person. This process of alternating race and Hispanic adjustments was repeated up to five more times ending with a race adjustment. The final value of PPSF for a person was the product of all of the PPSF(R)'s and PPSF(H)'s computed. The range of values of PPSF for each site is shown in the table below.

Table 5A: Range of Values of PPSF: 1996 & 1997

Site	1996	1997
Brevard, FL	0.8 - 1.9	0.8 - 2.2
Rockland, NY	0.6 - 1.8	0.6 - 1.9
Multnomah, OR	0.7 - 1.6	0.6 - 1.7
Fulton, PA	0.7 - 1.5	0.9 - 1.3
Douglas, NE		0.7 - 1.5
Otero, NM		0.5 - 2.4
Franklin, OH		0.6 - 1.4
Houston, TX		0.6 - 1.9

This completed the weighting of persons. The final person weight is shown below. The last step was to convert the final person to an integer value in a controlled rounding process that ensured that the rounded estimates of population by race, sex, or Hispanic origin were close to the unrounded estimate for that same block, tract, or county.

$$Final\ Person\ Wgt = BW\ x \ldots x\ HPF1\ x\ PPSF$$

6.0 Final Housing Unit Weighting Factors

6.1 Principal Person Factor (PPF)

This factor transferred some of the over or under-coverage detected in the person weighting onto the housing units. The PPSF factor of one of the householders of a housing unit was assigned to that housing unit. When assigned to a housing unit, this factor was renamed PPF. PPF for un-occupied housing units was 1.0.

6.2 Housing Post Stratification Factor #2 (HPF2)

Like HPF1, this factor made the number of housing units in a tract equal to the current MAF control count totals. It was computed in the same fashion as HPF1 except that the denominator term included all factors through PPF.

HPF2 was then computed in the same manner as HPF1 using the same MAF tract housing unit counts ($MAFHU_{ct}$) used for computing HPF1. The final HU weight was:

$$Final HU\ Wgt = BW\ x \ldots x\ HPF1 x\ PPF\ x\ HPF2$$

In a final step, the final housing unit weight (shown above), was converted to an integer value in a controlled rounding process. This process ensured that the rounded estimate of housing units within any individual block, tract, or county was within 1.0 of the unrounded estimate for that same block, tract, or county.

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.