Appendix B – 2019 NSCH Sample Frame and Sampling Flags Creation

2019 National Survey of Children’s Health sample frame

John Voorheis
Center for Economic Studies
Research and Applications
US Census Bureau
john.l.voorheis@census.gov
301-763-5326

April 16, 2018

This document describes using administrative records to build a sample frame for the National Survey of Children’s Health (NSCH) for 2019. We include tables and figures for the 2018 sample frame for reference.

Population of interest

The population of interest is all children residing in housing units in the US on the date of the survey.

A sample frame for all households with children

The sample frame identifies three mutually exclusive strata:

- [1] Households with *explicit links to children* in administrative data.
- [2a] Households without explicit links to children in administrative data, but predicted to be *likely to have children* conditional on administrative data.
- [2b] Households without explicit links to children in administrative data, but predicted to be *unlikely to have children* conditional on administrative data.

This document first explains the construction of the Stratum 1 flag, and then documents the separation of Strata 2a and 2b.

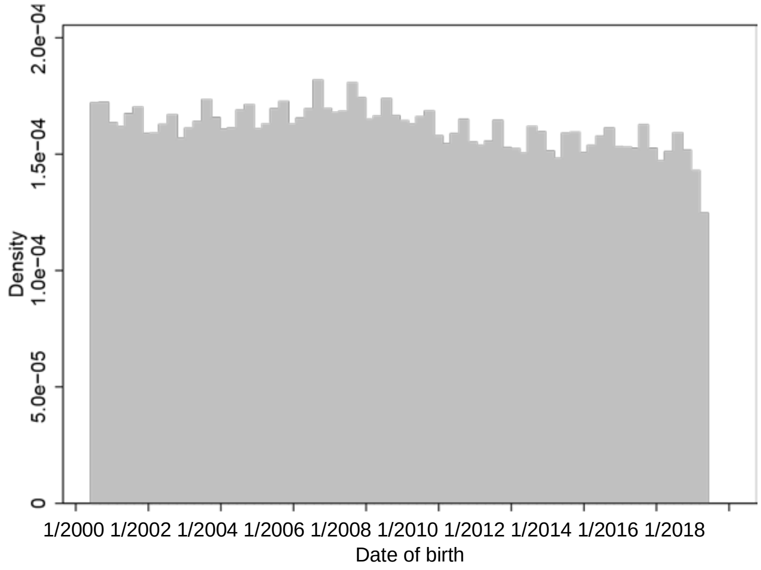
Stratum 1: identifying explicit links from children to addresses

The Stratum 1 flag for all households with explicit links to children comes from three data sources: the Numident, a list of Social Security Number applicants with data updated from various administrative records; and the CARRA kidlink file, a prototype linkage between children and parents based on Census and administrative records. Household addresses are updated with the Master Address Auxiliary Reference File, a file that links person identifiers with the latest location updates from a variety of administrative data.

Using the Numident to identify children

The Numident is based on off the all individuals who have been assigned Social Security Numbers. Demographic data from the Numident is updated from federal tax data and various administrative records. There are about 83 million children in the 2017 Numident who will be aged 0–17 years on June 1, 2018. Figure 1 shows the distribution of date of birth for these children.

Figure 1: Distribution of date of birth, aged 0–17 years as of June 1, 2018 (2017 Numident)



Identifying the households containing the children in the Numident

To sample households with children, we must connect the children in the Numident to the households in which they live. We do this with the CARRA kidlink file.

CARRA kidlink

The CARRA kidlink file uses data from Census survey and federal administrative records to link children PIKs to parent PIKs. We can use this file to identify the parents of children in the Numident.

The source data for the CARRA kidlink file are: the Census Numident, the 2010 Census Unedited File, the IRS 1040 and 1099 files, the Medicare Enrollment Database (MEDB), Indian Health Service database (IHS), Selective Service System (SSS), and Public and Indian Housing (PIC) and Tenant Rental Assistance Certification System (TRACS) data from the Department of Housing and Urban Development. Of these, the IRS 1040 provides the most significant information.

In the CARRA kidlink file generated March 2018, there are about 66 million unique records for children who will be aged 0–17 years on June 1, 2018.

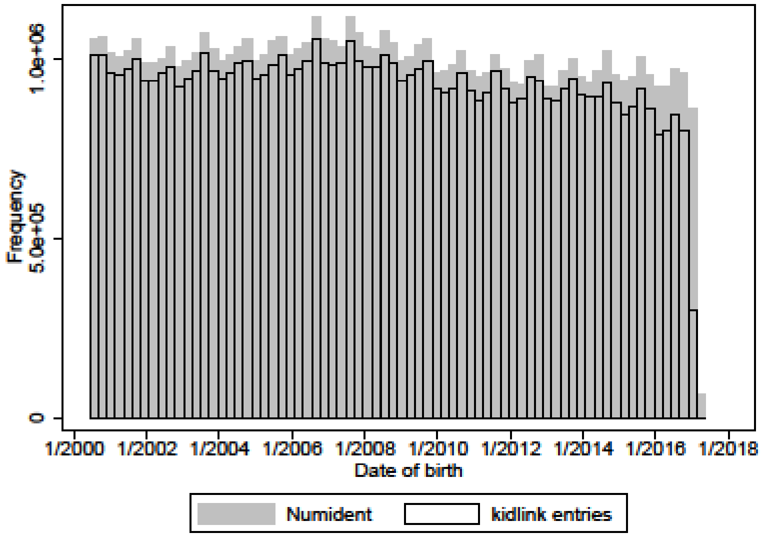
Let us consider how many children from the Numident have been linked to a parent in the CARRA kidlink file. Table 1 shows the number of children linked with both a mother and a father, linked with a mother only, linked with a father only, or not linked with any parent.

Table 1: Child-parent links in the CARRA kidlink file relative to the Numident population, aged 0–17 years as of 2018, March 2018 CARRA kidlink file

Type of link	Frequency	Percent
Mother and father	57,920,000	70%
Mother only	15,380,000	19%
Father only	2,836,000	3.4%
No link	6,821,000	8.2%
All children in Numident	82,956,000	100%

Figure 2 compares the distributions of date of birth for these children against the distribution shown in Figure 1.

Figure 2: Frequency distributions of date of birth, Numident vs. kidlink entries, aged 0–17 years as of June 1, 2018



The CARRA kidlink file was updated in March 2018 for NSCH sample frame production. We will use the same CARRA kidlink file for production in 2019. We will, however, supplement this file with additional parent-child linkages identified in sources which are not used to build the CARRA kidlink file, including ACS and CPS-ASEC data.

Updating household location using the MAF-ARF

In order to update household location, we use a Census dataset called the Master Address Auxiliary Reference File (MAF-ARF). The MAF-ARF links person identifiers to address identifiers using Census survey data and federal administrative data. The source data for the MAF-ARF file are: the Census Numident, the 2010 Census Unedited File, the IRS 1040 and 1099 files, the Medicare Enrollment Database (MEDB), Indian Health Service database (IHS), Selective Service System (SSS), and Public and Indian Housing (PIC) and Tenant Rental Assistance Certification System (TRACS) data from the Department of Housing and Urban Development, and National Change of Address data from the US Postal Service. Of these, the IRS 1040 provides the most significant information.

Out of about 83 million children in the Numident, about 68 million, are matched directly to a MAFID. Out of about 73 million kidlink-matched mothers, about 67 million are matched to a MAFID. Out of about 60 million kidlink-matched fathers, about 56 million are matched to a MAFID.

For each child observation from the Numident, we now have three possible MAFIDs: the kid to MAF-ARF MAFID, the child-to-kidlink-to-mother-to-MAF-ARF MAFID, and the child-to-kidlinkto-father-to-MAF-ARF MAFID. I allocate the single MAFID using that order. First, I assign the directly identified child MAFID (about 65 million cases). If the MAFID is missing, I assign the mother MAFID (about 6 million cases). Finally, if the MAFID is still missing, I assign the father MAFID (about 3 million cases). That leaves about 9 million children from the Numident not assigned MAFIDs (a MAFID match rate of 87.2%).

There are some MAFIDs associated with a great number of children. As an example, out of 74 million children associated with a MAFID, about 7 million children are associated with a MAFID with more than 20 child-MAFID links.

The 74 million children associated with a MAFID are then collapsed down to 38 million unique MAFIDS. This implies 1.94 children per household for households assigned a flag.

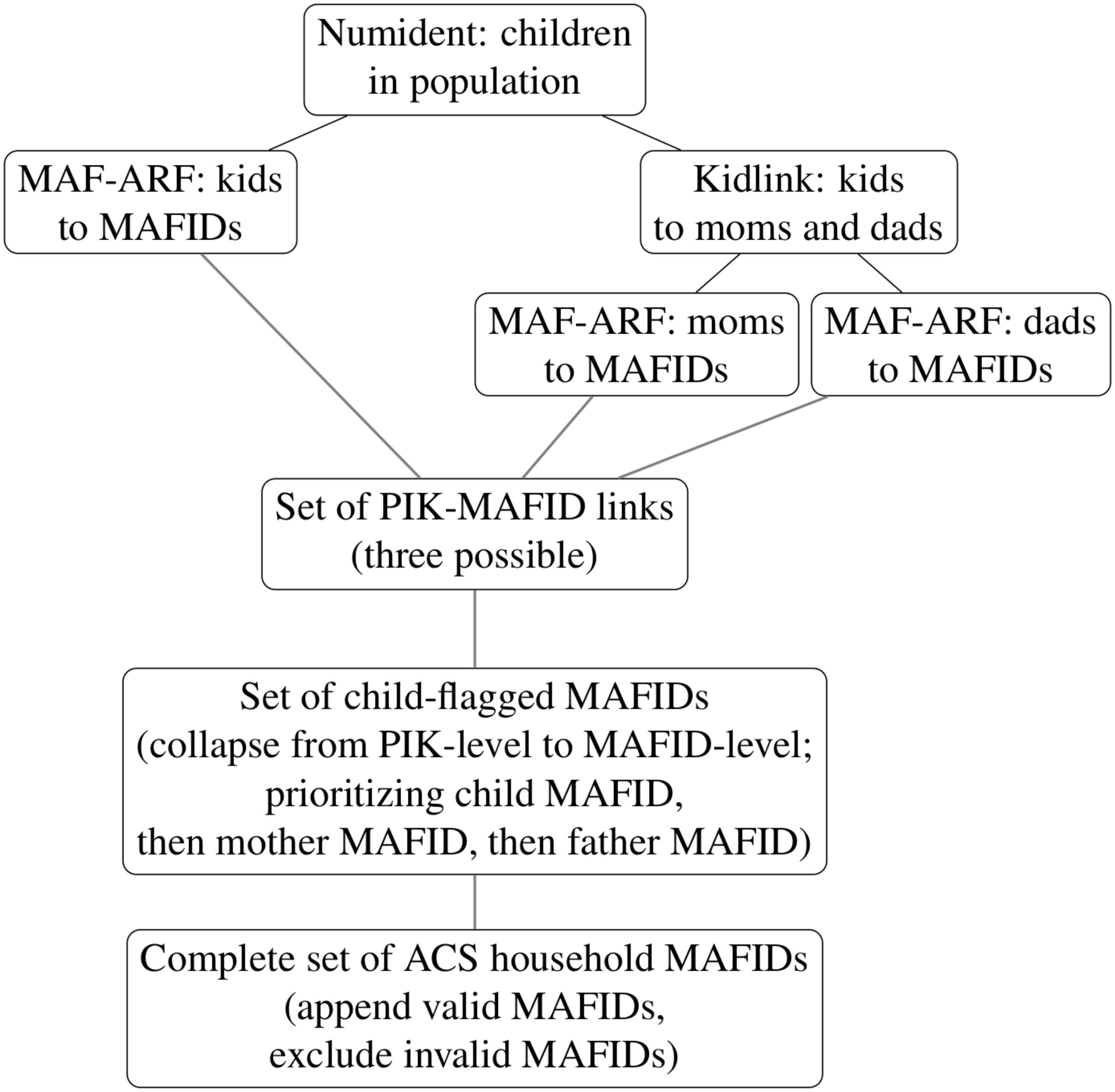
For 2019, one additional step will be conducted in the construction of stratum 1. We will use administrative HUD PIC and TRACS data, which contain flags for the number of children present at the household level for all public housing and voucher households, to enhance the existing stratum 1 process. We will merge all MAFIDs not assigned a stratum 1 flag using the above kidlink-MAF-ARF process, with the most recent data on all public housing and voucher households in the PIC-TRACS data. We will then assign a stratum 1 flag to all households which have a child present flag in the HUD data.

We then need to scale up the MAFID list to the universe of MAFIDs to allow sampling of unflagged households. A merge of the 38 million unique child-flagged MAFIDS with the January 2018 ACS MAF-X file matches 38 million MAFIDS with child flags, removes 164 million MAFIDS with child flags, and adds 289 MAFIDs without child flags. The sample frame file now has about 203 million valid MAFIDS, of which 38 million MAFIDS include child flags. Compare this with the 2011 ACS, in which about 37 million out of 115 million households included related children.[1]

Stratum 1 construction visualization

Figure 3 shows a visualization of the sample frame construction.

Figure 3: Stratum 1 construction


Numident: children in population
MAF-ARF: kids to MAFIDs
Kidlink: kids to moms and dads
MAF-ARF: moms to MAFIDs
MAF-ARF: dads to MAFIDs
Set of PIK-MAFID links (three possible)
Set of child-flagged MAFIDs (collapse from PIK-level to MAFID-level; prioritizing child MAFID, then mother MAFID, then father MAFID)
Complete set of ACS household MAFIDs (append valid MAFIDs, exclude invalid MAFIDs)

[1] http://www.census.gov/prod/2013pubs/p20-570.pdf

Strata 2a and 2b: identifying probabilistic links from children to addresses

In 2016, the Stratum 1 flag performed well. That is, it contained approximately the same rate of children after as sampling as had been predicted before the survey. The survey team would like to further increase the sampling efficiency of the survey by adding more information to the second stratum. By definition, Stratum 2 does not have explicit links from children to households in the administrative data. In 2018 as in 2017, we will further bifurcate Stratum 2 into those households more likely to have children and those households less likely to have children.

Households will be assigned to Stratum 2a based on a model of child presence as a function of variables available in administrative data for all households in the MAF. The model is estimated with data from the most recent year of the ACS, in which child presence can be observed. Then parameter estimates from that model can be used to predict the likelihood of child presence for all households. These models are estimated separately for each state, and the threshold for bifurcation is based on an objective of minimizing the size of Stratum 2a while also maintaining 95% coverage of children in Strata 1 and 2a.

Definitions

Population or sample concepts

- 2016 ACS sample, edited and swapped
  - unit of observation is the household, unless noted otherwise
  - sample includes sampled vacant dwellings, unless noted otherwise
- MAF
  - population but restricted to MAFIDs marked as valid for ACS

Sample frame notation

- h indexes household
- s indexes states
- C equals 1 if a household has any children, 0 otherwise
- Strata:
  - S_1: household with children
  - S_{2a}: household likely to have children – S_{2b}: household unlikely to have children
- Strata sizes:

  - $p(S_1)$
  - $p(S_{2a})$
  - $p(S_{2b})$

- Strata child rates:
  - $p(C|S_1)$
  - $p(C|S_{2a})$
  - $p(C|S_{2b})$

- Coverage with unsampled S_{2b}:
  - $p(S_1 \cup S_{2a}|C)$

Model

Our goal is a scalar measure of the likelihood of a child being associated with a MAFID. This measure must be available for all ACS-valid MAFIDs in the MAF. Using a sample in which the presence of children is observable, we will estimate a model of child presence. The regressors used to make the index prediction must be observable for all MAFIDs (i.e., to predict outside of the estimation sample to the entire MAF).

The general model is:

$$C_h = f(X_h;\theta),$$

where C is equal to one if a household includes any children and zero otherwise, X is a vector of characteristics available for all households, and θ is an unknown vector of parameters.

We estimate the model using the most recent ACS 1-year sample:

$$E[C_h|X_h] = f(X_h;\hat{\beta}_{ACS}) \text{ for households } h \text{ in the ACS.}$$

With parameter estimates from the ACS, we make predictions for the entire MAF:

$$\hat{C}_h = f(X_h;\hat{\beta}_{ACS}) \text{ for households } h \text{ in the MAF.}$$

In practice, we estimate models separately for each state. We do this to account for systematic differences in administrative records coverage and MAF quality across states. The model can now be specified as:

$$E[C_{hs}|X_{hs}] = f(X_{hs};\hat{\beta}_{s,ACS}) \text{ for households } h \text{ in state } s \text{ in the ACS,}$$

where s is the MAFID's state and the parameters $\hat{\beta}_{s,ACS}$ now vary across states. The state-specific predictions become:

$\hat{C}_{hs} = f(X_{hs};\hat{\beta}_{s,ACS})$ for households h in state s in the MAF.

Estimation

The model above is estimated as a linear probability model separately for each state using the edited and swapped 2015 ACS sample. The outcome is child_present, a flag for whether a child is present at the sampled MAFID.

The following covariates are included (with associated data sources) and are available for each MAFID (except where a missingness flag is used):

- 2016 ACS 5-year published aggregate data
  - acs_blkgrp_childrate_lvout: proportion of residents of block group who are children, excluding the own-observation child counts from the numerator and denominator
- MAF-ARF
  - female2050: flag for female between ages 20 and 50 at MAFID
  - adult2050: flag for adults between ages 20 and 50 at MAFID
  - coresid_sexdiff: flag for coresidence of men and women between ages 20 and 50 at MAFID
  - miss_adult2050: flag for missingness from MAF-ARF
- IRS 1040 filings, tax year 2015
  - any_kid_deduct_max: does any tax form associated with this MAFID have any deduction related to children?[2]
  - itemized_max: does any tax form associated with this MAFID use itemized deductions?
  - miss_any_kid_deduct_max: flag for MAFIDs without associated tax forms
- VSGI NAR commercial data
  - vsgi_nar_homeowner_max: does any observation associated with this MAFID record it as homeowener-occupied?
  - miss_vsgi_nar_homeowner_max: flag for MAFIDs without associated VSGI data
- Targus commercial data
  - targus_homeowner_0: various flags for homeowner-occupied MAFID
  - targus_homeowner_A: various flags for homeowner-occupied MAFID
  - targus_homeowner_B: various flags for homeowner-occupied MAFID
  - targus_homeowner_C: various flags for homeowner-occupied MAFID

[2] The following IRS variable were used to make this variable: child exemptions and EITC qualifying children.

- targus_homeowner_D: various flags for homeowner-occupied MAFID
- targus_homeowner_E: various flags for homeowner-occupied MAFID
- targus_homeowner_F: various flags for homeowner-occupied MAFID – miss_targus_homeowner: flag for MAFIDs without associated Targus data

Parameter estimates are stored in the file frame2018_child_present_bystate.csv.

Sample frame objective function

In order to choose an optimal Strata 2a, we use the following objective function:

- Minimize the size of Strata 2a while maintaining coverage of at least 95%

Strata 2a is defined as:

$$S_{2a} = \{\text{households in the MAF with } \hat{C}_h > \bar{C} \text{ but not in } S_1\}.$$

Strata 2b is defined as

$$S_{2b} = \{\text{households in the MAF but not in } S_1 \text{ or } S_{2a}\}.$$

With state-specific modeling, the objective function and coverage constraint also becomes state specific:

- Minimize the size of Strata 2a in each state while maintaining coverage of at least 95% in each state

State-specific Strata 2a is defined as:

$$S_{2a} = \{\text{households in the MAF with } \hat{C}_{hs} > \bar{C}_s \text{ but not in } S_1\}.$$

Strata 2b is defined as before.

Optimization algorithm

The optimization parameter is a threshold on the child-present prediction probability, such that MAFIDs with values above the threshold are assigned to Stratum 2a. Starting at a low threshold ($\bar{C}$)[3], follow this algorithm:

[3] The most conservative starting threshold would be at $p(S_1)$, where $p(S_{2b}) = 0$.

1. Under the current threshold $\bar{C}$, calculate the proportion of MAFIDs in Stratum 2a, $p(S_{2a})$, and the coverage of Strata 1 and 2a under no sampling of Strata 2b, ($p(S_1 \cup S_{2a}|C)$).
2. If $p(S_{2a}) > 0$ and $p(S_1 \cup S_{2a}|C) \geq 0.95$, then increase the child prediction threshold $\bar{C}$ one step (e.g., 0.01) and return to (1). If $p(S_1 \cup S_{2a}|C) < 0.95$, then the previous threshold $\bar{C}$ is the optimal cutoff for S_{2a}.

Under state-specific modeling, this algorithm is applied separately to each state.

Optimal strata

Table 2 shows the optimal strata under a 95% coverage constraint for Strata 1 and 2a. The coverage constraint assumes non-sampling of Stratum 2b. The notation is as defined above. The strata were optimized separately for each state using parameter estimates from separate state regressions of child presence in the 2016 ACS microdata.

Auditing the sample frame against the ACS

To examine the performance of the administrative records used to build the sample frame, we merge the list of MAFIDs constructed above with the American Community Survey housing-unit sample from 2016. Currently, this audit uses unedited ACS data (i.e., item nonresponse are left as missing and are not imputed including children's age). If item nonresponse is random with respect to the presence of children in the household, this should not cause any systematic bias in the audit.

All estimates are weighted with the housing-unit-level weights, which include weight for vacant units (209,556 vacant housing units in the 2016 ACS). In vacant housing units, we assign zero children. These estimates should reflect the NSCH survey production process.

State-specific performance

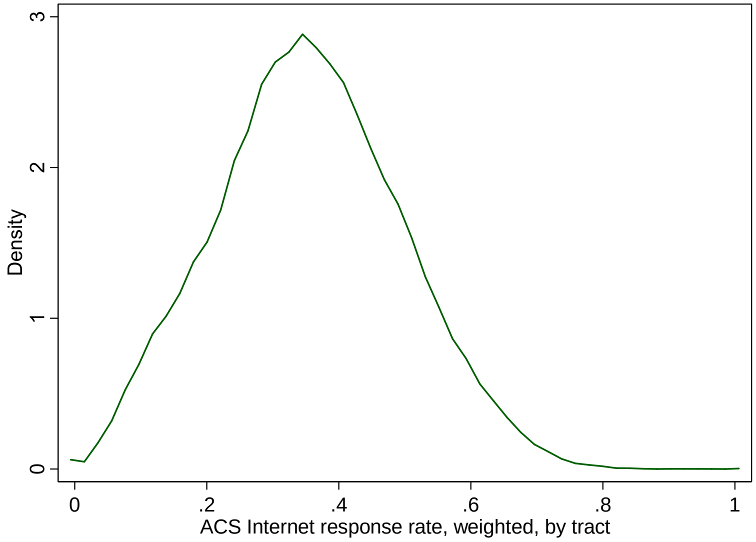
In 2018, the smallest oversample strata were in Hawaii, Maine, Vermont, and West Virginia. The largest oversample strata are in California, Texas, and Utah. The highest rates of Type 1 error are in DC, Florida, Louisiana, Mississippi, Nevada, and South Carolina. The highest rates of Type 2 error were in Alaska, Hawaii, New Mexico, Texas, and Utah. For 2019, we will perform similar audits of the frame against the 2017 ACS , and will additionally audit the frame against an early release file of 2018 ACS microdata.

Local-area Internet-accessibility

Here we describe the construction of a tract-varying Internet-accessible household flag.

Since 2012, ACS respondents have been able to submit survey forms over the Internet. ACS paradata record whether a respondent chose the online option. The ACS paradata has been summarized at the tract level. Our Internet-accessible household measure is equal to a weighted proportion of the respondents that chose to submit the ACS survey over the Internet if given the option to do so. Figure 4 shows the kernel-smoothed distribution of tract-level Internet response for the 2013–2014 ACS survey years.

Figure 4: Kernel-smoothed probability distribution function of tract-level ACS Internet response rate, ACS paradata, 2013–2014 survey years

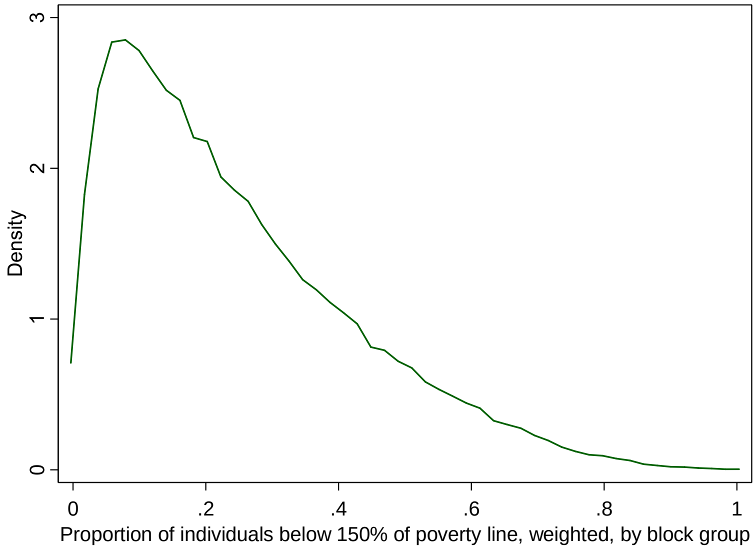


To construct an Internet-access flag, we use the first tritile for a cut-off. A block is considered to have low Internet access if the Internet accessibility index is below the first tritile of the block-level distribution. For low-population blocks, we replace missing values of the block-varying low-Internet flag with the modal value from the corresponding block group. For very new housing units without assigned Census blocks, we assign a value of zero for this binary variable (i.e., the default for these new households is high Internet accessibility.)

Local-area household income relative to the poverty rate

The frame has a set of poverty variables from the 2016 5-year American Community Survey file. These variables measure the proportion of households with household income in an interval defined by the poverty rate. Figure 5 shows the kernel-smoothed probability distribution function of the proportion of households in the block group that have household income less than 150% of the poverty rate.

Figure 5: Kernel-smoothed probability distribution function of block-group-level 150% poverty rate, ACS, 2016 5-year file



Final sample frame data layout

The component data files are merged together based on MAFID. The data layout for this combined file is given in Table 2.

Table 2: NSCH population data file layout

Variable name	Label	Level of variation	Type	Domain	Any missing?
mafid	Master Address File ID	MAFID	long	9 digits	no
maf_curstate	State	State	str2		no
maf_curcounty	County	County	str3		no
maf_curblktract	Tract	Tract	str6		yes

maf_curblkgrp	Block group	Block group	str1		yes
maf_curblk	Block	Block	str4		yes
stratum1	Stratum 1 identifier	MAFID	byte	{0,1}	no
stratum2a	Stratum 2a identifier	MAFID	byte	{0,1}	no
stratum2b	Stratum 2b identifier	MAFID	byte	{0,1}	no
acs_tract_net_response	ACS Internet response	Tract	float	[0,1]	yes
web_low	Low web use (lowest tritile)	Tract	byte	0,1	no
blkgrp_lt_100_povrate	Pr. HH w/ inc. < 100% poverty rate	Block group	float	[0,1]	yes
blkgrp_100_150_povrate	Pr. HH w/ inc. 100–150% poverty rate	Block group	float	[0,1]	yes
blkgrp_150_185_povrate	Pr. HH w/ inc. 150–185% poverty rate	Block group	float	[0,1]	yes
blkgrp_185_200_povrate	Pr. HH w/ inc. 185–200% poverty rate	Block group	float	[0,1]	yes
blkgrp_gt_200_povrate	Pr. HH w/ inc. > 200% poverty rate	Block group	float	[0,1]	yes
blkgrp_lt_150_povrate	Pr. HH w/ inc. < 150% poverty rate	Block group	float	[0,1]	yes
valdf18	Valid mailing address	MAFID	byte	{0,1}	yes

Filename: nsch_pop_file.sas7bdat
Population: all MAFIDs in 2017 MAF-X
Unit of observation: household (MAFID)
Number of observations: 202,800,000
Filesize: 20GB