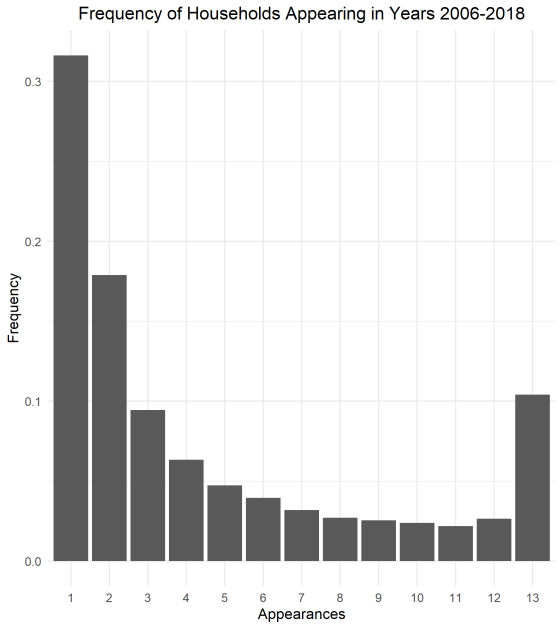
North Carolina InfoUSA Analysis

Davis Berlind

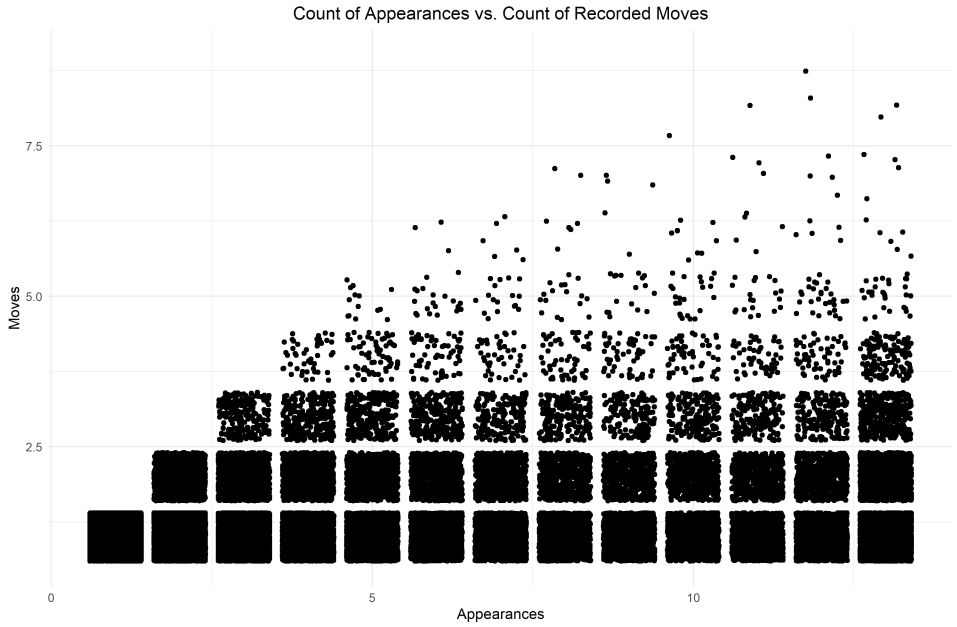
November 2019

1 Migration

In the combined InfoUSA data for North Carolina (2006-2018), there are 13,874,976 households observed (as indicated by unique values for `FAMILYID`, which remains stable across the years of the data set). The average household appears in the North Carolina subset of the data 4.5 times (sd 4.12). The distribution of appearances in the data set is displayed below.



We can indicate the number of times a household moves by checking how many unique street addresses are associated with that household in the data. Below we plot the number of moves against the number of times the household appears in the data set. Even for households that appear in nearly all years, the number of moves is concentrated around the lower end of the distribution of move counts.



To understand the determinants of a household appearing in the data set, we regress the count of appearances on household/head of household characteristics. For households appearing at least four times, we also regress the count of moves on household/head of household characteristics. The results below tell us that the head of the household's age is the main determinant for whether a household will included in more years of the InfoUSA data. A household where the head is 65-69 years old has three more years of available data than a household where the head is under 25 years old, *ceteris paribus*. We also see that married households, households with children, households that are owned, and wealthier households appear in the data more times (though each of these variables are also very correlated with head of household age). The result of the regression on move counts indicates that those near the middle and very end of the age distribution are more likely to move, as are black and wealthier residents. We also have the intuitive result that the number of moves decreases as the length of residence and likelihood of being a homeowner increase.

	Appearances		Moves	
	Estimate	SE	Estimate	SE
Intercept	-1.31	(0.008)***	0.31	(0.005)***
Mean Log HH Income	0.523	(0.002)***	0.11	(0.001)***
Mean Own/Rent Status	0.13	(0.001)***	-0.07	(0.0002)***
Mean Children	1.47	(0.002)***	0.02	(0.001)***
Mean Marital Status	0.12	(0.001)***	0.01	0.003)***
Mean Length Residence	0.19	(0.0001)***	- 0.01	(0.000)***
Race				
Black	0.62	(0.007)***	0.11	(0.003)***
Hispanic	0.15	(0.007)***	-0.01	(0.003)***
Mixed	-0.63	(0.073)***	0.04	(0.027)
Native American	0.19	(0.029)***	0.04	(0.009)***
Pacific Islander	0.37	(0.091)***	0.07	(0.033)*
White	0.27	(0.006)***	0.06	(0.002)***
Other	-0.08	(0.008)***	0.04	(0.003)***
Head HH Age				
Age 25-29	0.15	(0.005)***	0.19	(0.004)***
Age 30-34	0.48	(0.005)***	0.27	(0.004)***
Age 35-39	0.84	(0.005)***	0.30	(0.004)***
Age 40-44	1.11	(0.005)***	0.25	(0.004)***
Age 45-50	1.58	(0.005)***	0.19	(0.004)***
Age 50-54	1.85	(0.005)***	0.15	(0.004)***
Age 55-59	2.26	(0.005)***	0.14	(0.004)***
Age 60-54	2.64	(0.005)***	0.13	(0.004)***
Age 65+	2.84	(0.006)***	0.03	(0.004)***
Age 65-69	3.05	(0.006)***	0.19	(0.004)***
Age 70-74	3.23	(0.006)***	0.21	(0.004)***
Age 75+	2.99	(0.005)***	0.28	(0.004)***
R^2	0.524		0.101	

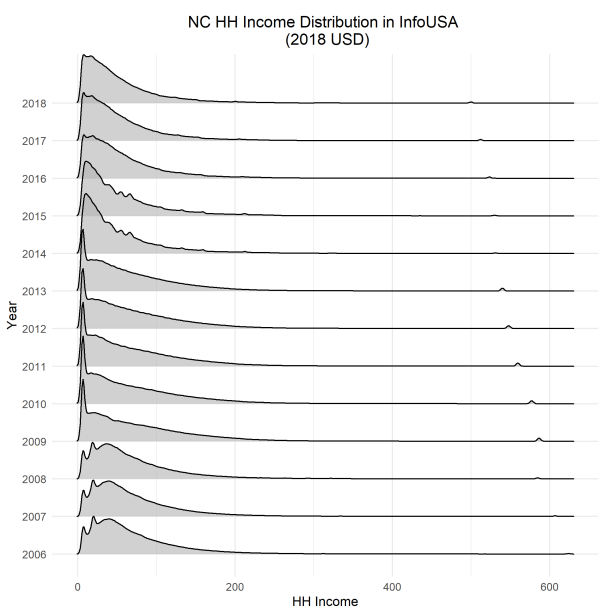
2 Income

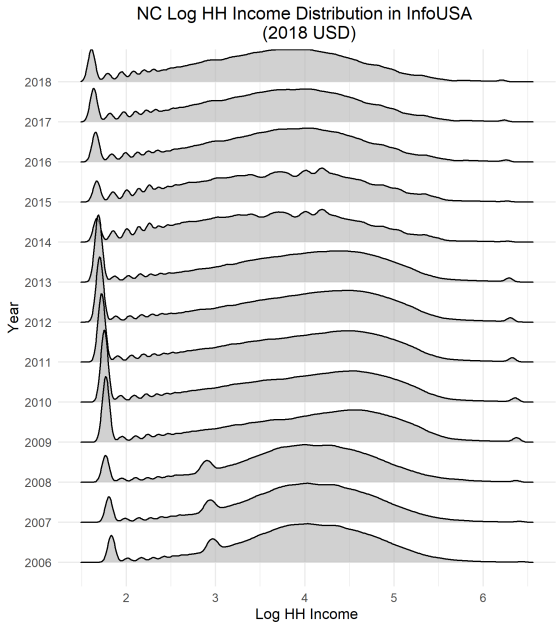
The first table below displays summary statistics for the household income distribution of the North Carolina subset of InfoUSA. Note that household incomes are reported to the nearest $1,000 and are filtered at $5,000 and $500,000 on the lower and upper ends respectively. Notice how incomes are increasing, up until 2009 when they peak, and are decreasing there after with the onset of the recession.

Year	Mean	Median	SD	Mean (log)	Median (log)	SD (log)
2006	69.9	54.9	58.8	3.93	4.01	0.843
2007	70.3	54.6	60.1	3.94	4.00	0.843
2008	71.5	54.9	64.8	3.93	4.01	0.872
2009	79.4	58.7	79.3	3.90	4.07	1.07
2010	75.1	54.2	76.5	3.83	3.99	1.09
2011	73.3	53.7	74.2	3.81	3.98	1.08
2012	74.0	54.8	73.9	3.83	4.00	1.07
2013	70.1	50.7	71.0	3.77	3.93	1.06
2014	54.9	37.2	55.3	3.58	3.62	0.946
2015	57.8	40.3	57.2	3.64	3.70	0.942
2016	60.4	43.9	60.3	3.70	3.78	0.938
2017	57.5	41.0	58.0	3.64	3.71	0.949
2018	55.9	40	57.0	3.61	3.69	0.948

Table 1: Income Summary Statistics for InfoUSA (1,000 2018 USD)

The plots below display the full income distributions for North Carolina in the years 2006-2018. Notice that there is quite a bit of clustering around the lower threshold (the left hand spike in the log plots), and in the years 2006-2008 there is an additional cluster of incomes between the mean and left tail.





We now consider a comparison between household incomes in InfoUSA and the 5-Year ACS (i.e. the 5-Year PUMS data made publicly available by the Census Bureau). We consider two periods of data, 2006-2010 and 2012-2016. Note that the incomes in the ACS have been transformed to have the same upper and lower bounds as in InfoUSA. We have also omitted households classified as group quarters. We provide both raw and weighted estimates from the ACS, the latter of which are calculated using the sample weights provided within the PUMS data.

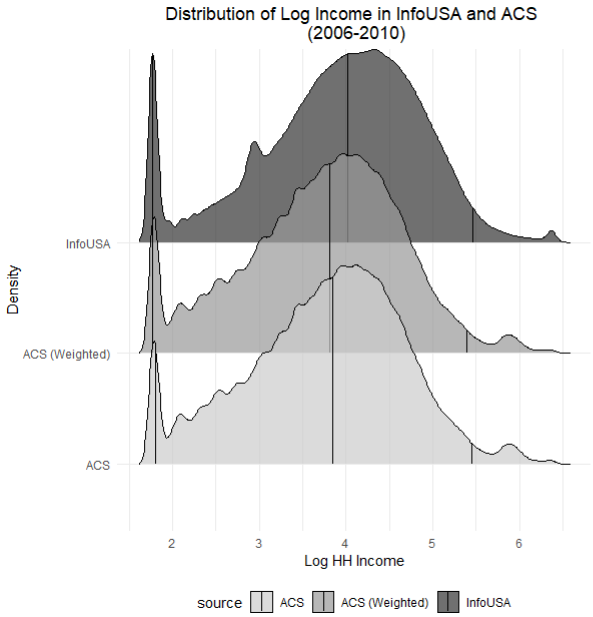
Period	Mean			Median			SD		
	InfoUSA	ACS	ACS Weighted	InfoUSA	ACS	ACS Weighted	InfoUSA	ACS	ACS Weighted
2006-2010	73.3	63.7	61.3	55.4	46.8	45.0	68.7	64.4	61.9
2012-2016	63.3	69.2	66.8	44.6	49.5	47.9	64.3	72.5	69.7

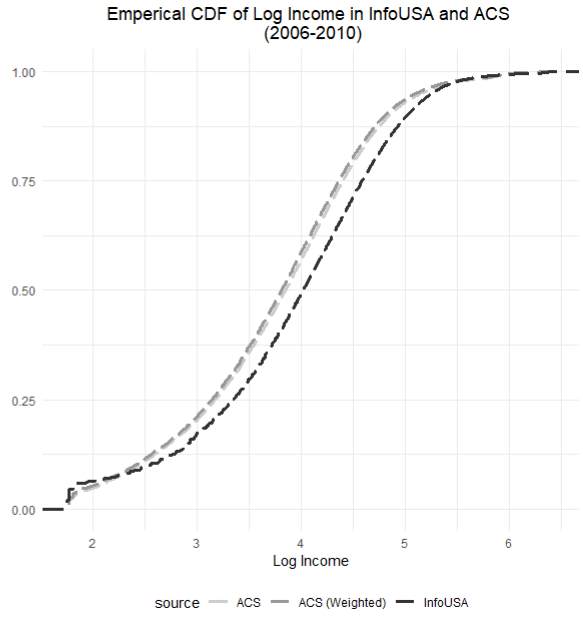
Table 2: Household Income in InfoUSA vs. ACS (1,000 2018 USD)

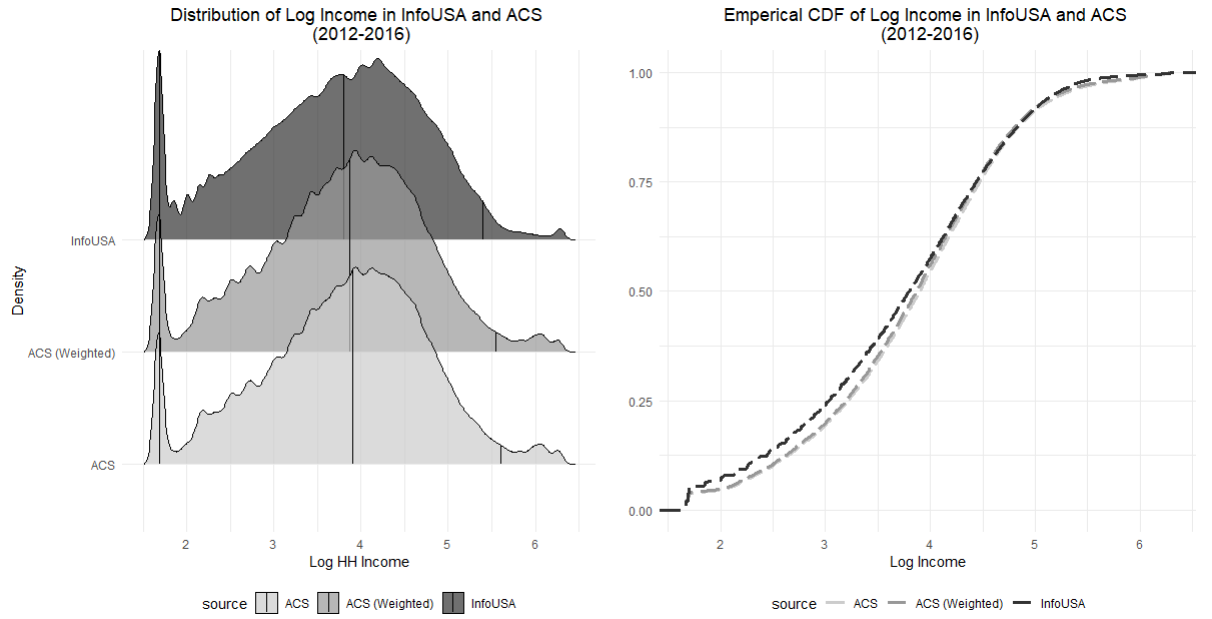
Period	Mean			Median			SD		
	InfoUSA	ACS	ACS Weighted	InfoUSA	ACS	ACS Weighted	InfoUSA	ACS	ACS Weighted
2006-2010	3.90	3.76	3.72	4.02	3.86	3.72	0.953	0.921	0.917
2012-2016	3.70	3.82	3.78	3.80	3.90	3.87	0.996	0.954	0.947

Table 3: Log Household Income in InfoUSA vs. ACS (Log 1,000 2018 USD)

The plots below display the full distributions of log household income for each period as well as the empirical CDFs. For the period 2012-2016, a visual check does not give an indication of differences betweent the distributions.







3 Race

In the next table we examine the distribution of the head of household's race. Note that in the earlier periods of the InfoUSA data, a large proportion of the observations are missing data on race. To compensate for this we use the `wru` package in `R`, which implements the methods described in Imai, K. and Khanna, K. (2016) (the function we rely on uses the Census Bureau's Surname List in combination with geocoded voter registration records to conduct Bayesian posterior inference on race). The results of the actual and estimated distributions are reported below, as are the race distributions for North Carolina from the ACS.[1]

	Hispanic			White			Black			Asian			Other		
Period	InfoUSA	InfoUSA (wru)	ACS	InfoUSA	InfoUSA (wru)	ACS	InfoUSA	InfoUSA (wru)	ACS	InfoUSA	InfoUSA (wru)	ACS	InfoUSA	InfoUSA (wru)	ACS
2006-2010	2.0%	3.4%	7.8%	54.3%	77.2%	66.1%	5.6%	8.3%	21.2%	1.0%	1.2%	2.1%	37.1%	9.9%	2.8%
2007-2011	2.3%	3.4%	8.1%	57.8%	77.2%	65.7%	5.9%	8.2%	21.2%	1.0%	1.2%	2.1%	32.9%	10.0%	2.9%
2008-2012	2.5%	3.3%	8.3%	62.2%	77.3%	65.2%	6.5%	8.1%	21.2%	1.2%	1.3%	2.2%	27.6%	10.0%	3.0%
2009-2013	3.0%	3.3%	8.5%	67.6%	77.4%	64.9%	7.3%	8.2%	21.1%	1.4%	1.4%	2.4%	20.7%	9.8%	3.1%
2010-2014	3.4%	3.4%	8.7%	72.5%	77.6%	64.6%	8.1%	8.2%	21.2%	1.5%	1.4%	2.5%	14.6%	9.4%	3.2%
2011-2015	3.7%	3.6%	8.8%	76.0%	79.2%	64.2%	8.6%	8.3%	21.2%	1.7%	1.5%	2.6%	10.1%	7.5%	3.2%
2012-2016	4.0%	3.7%	8.9%	78.7%	80.7%	64.0%	9.0%	8.5%	21.2%	1.8%	1.5%	2.6%	6.6%	5.6%	3.3%
2013-2017	4.3%	3.9%	9.1%	80.0%	82.1%	63.6%	9.3%	8.6%	21.2%	1.9%	1.6%	2.8%	4.5%	3.7%	3.4%

Table 4: Racial Make Up of North Carolina in InfoUSA and ACS

Note that InfoUSA consistently under-represents black heads of households and over-represents white heads of households relative to the ACS. This trend is stable in every year, regardless of whether we look at the raw data or the estimates produced with `wru`, and it is consistent with what we see in the full InfoUSA data set.

[1]Note we have collapsed Pacific Islander/Native Hawaiian into the Asian category, and Mixed Race and Native American into the other category.

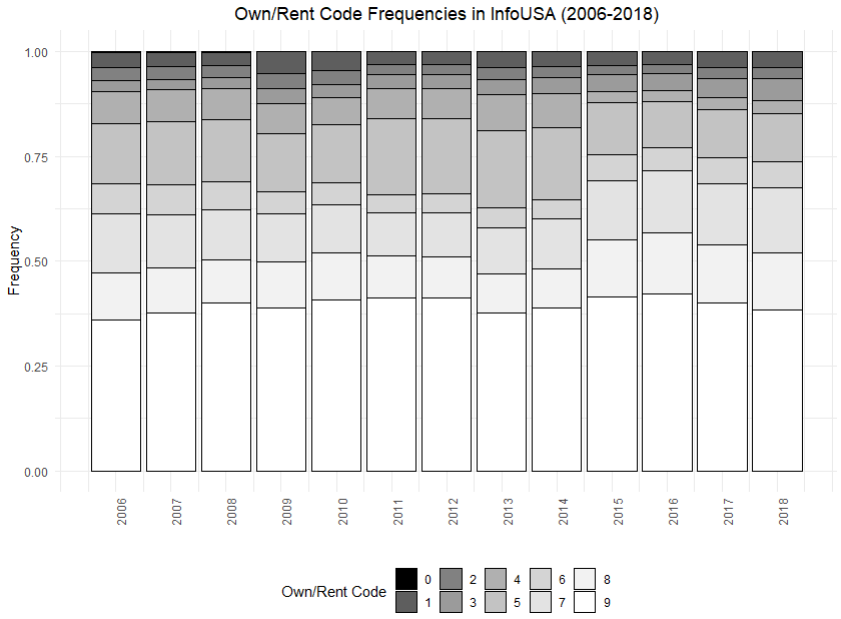
4 Tenure

The tenure of households in the InfoUSA data set is classified on a 0-9 scale with 0 and 9 representing confirmed rented and owned households respectively. The overall distribution of codes is as follows (note the large concentration on owners).

Code	Description	Proportion
0	Renter	0.09%
1	Likely Renter	3.59%
2	Likely Renter	2.79%
3	Unknown	3.6%
4	Unknown	5.96%
5	Unknown	14.52%
6	Likely Owner	5.61%
7	Likely Owner	12.78%
8	Likely Owner	11.52%
9	Owner	39.54%

Table 5: Own/Rent Code Summary (2006-2018)

The following plot shows how these proportions are changing over time in the data set.



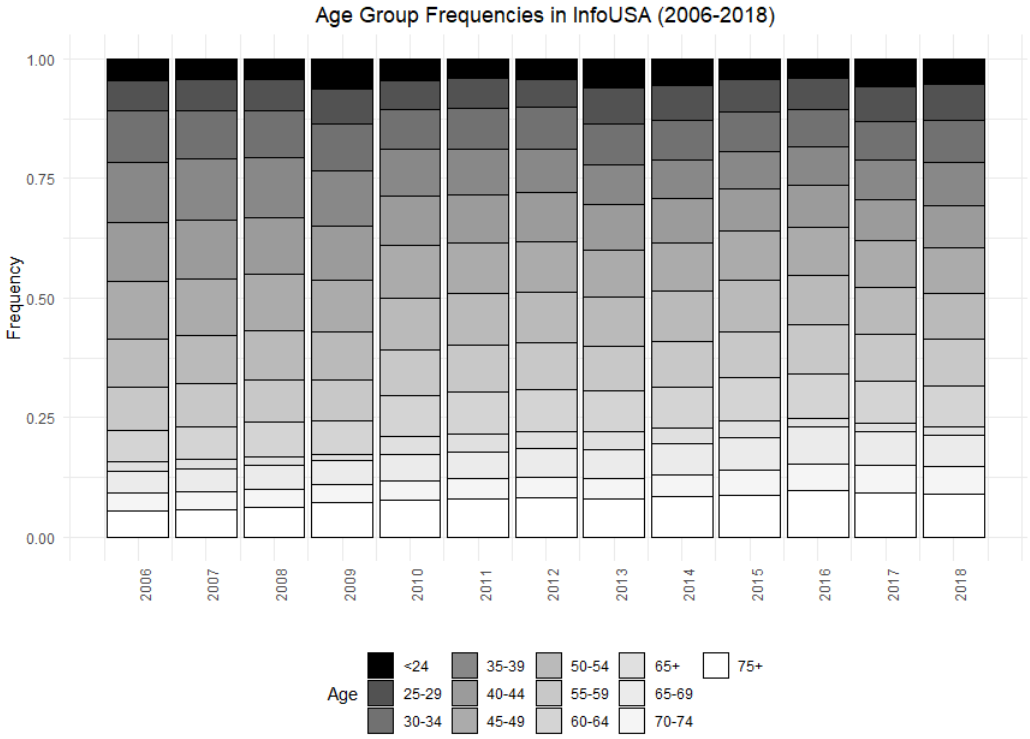
We can also calculate the overall percentage of owners and compare these with the estimates reported in the ACS. We estimate the share of owned and rented households two ways: first, by leaving out households coded as unknown, and second, by coding the unknown households as rented, which produces estimates more in-line with what is observed in the ACS.

Period	InfoUSA (Rent: 0-3, Own: 7-9)	InfoUSA (Rent: 0-6, Own: 7-9)	ACS
2006-2010	85.9%	61.9%	68.1%
2007-2011	86.0%	61.9%	67.8%
2008-2012	86.2%	62.0%	67.1%
2009-2013	85.6%	61.1%	66.4%
2010-2014	86.1%	60.9%	65.8%
2011-2015	86.7%	62.1%	65.1%
2012-2016	87.0%	64.1%	64.8%
2013-2017	86.7%	65.5%	65.0%

Table 6: Proportion of Owned Households in InfoUSA vs. ACS

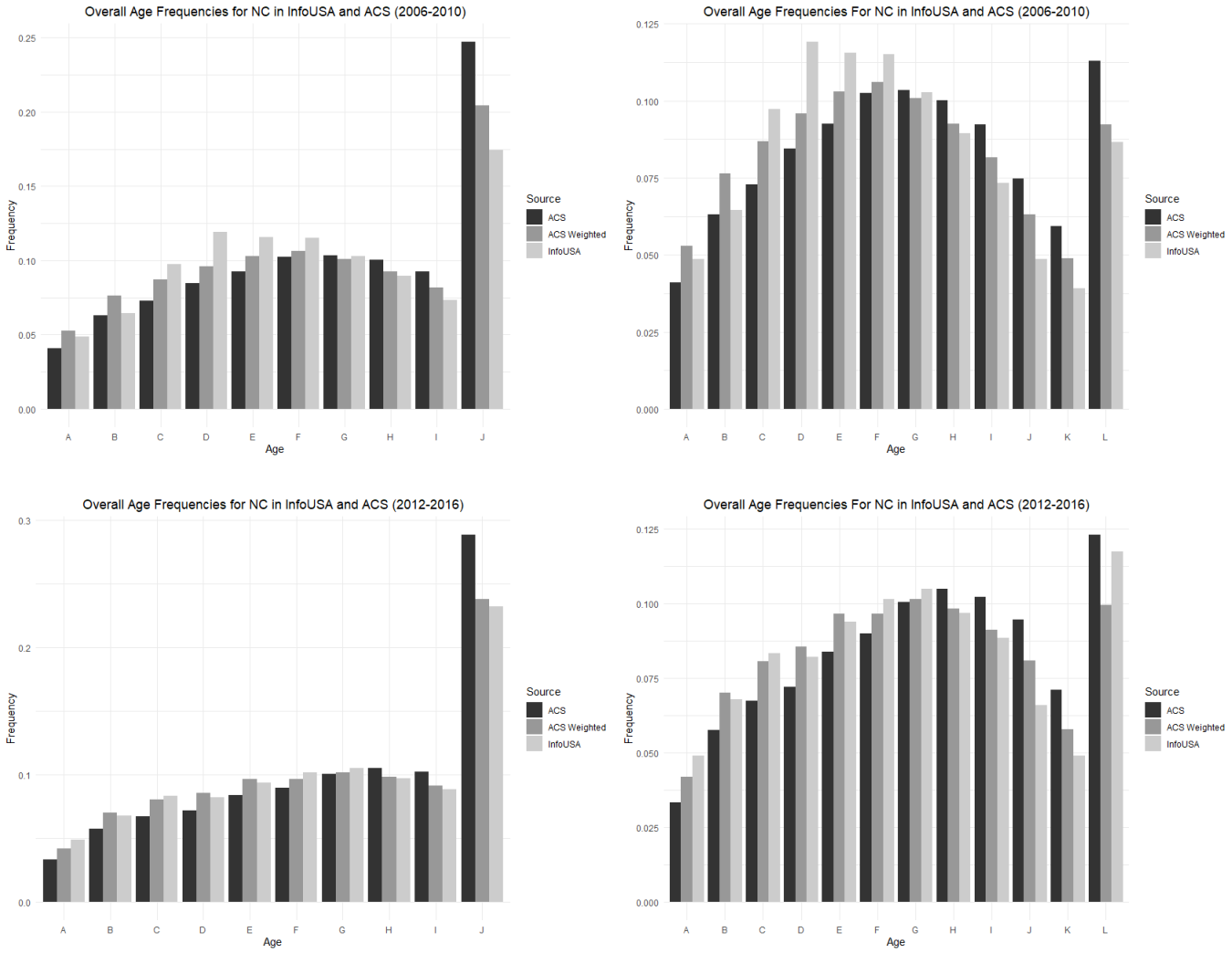
5 Age

We are given the range of the head of household's age in the InfoUSA data. The overall distributions of these ages for each year are plotted below.

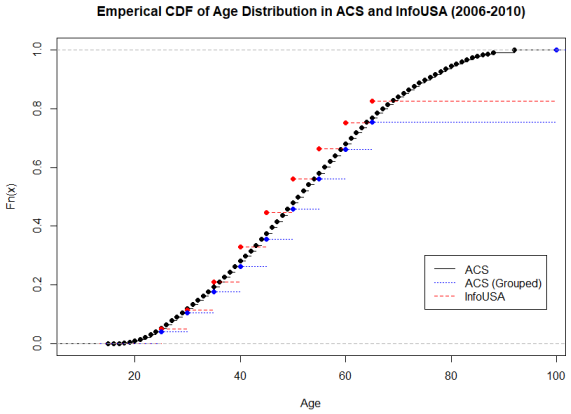


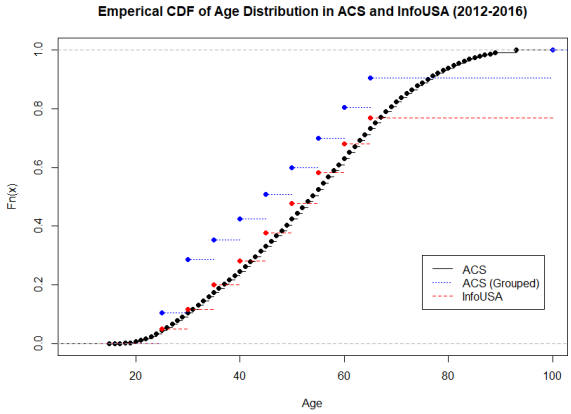
One tricky feature of this data is that we have a group for age 65+ as well as the age groups 65-69, 70-74, and 75+. The plots below show the distributions over the ages of heads of households in 2006-2010 and

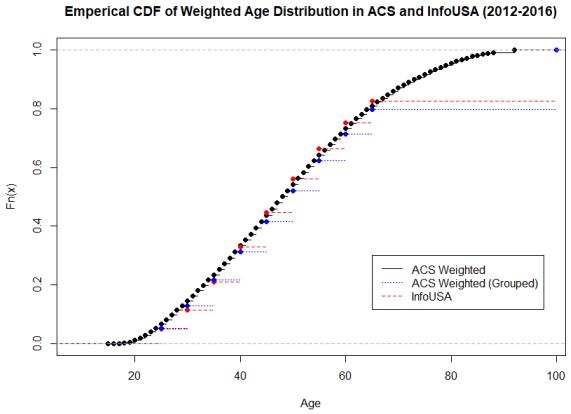
2012-2016 for the InfoUSA and PUMS data. In one set of plots we group all heads of household 65 years and older together into a single group, and in the other we choose to combine the 65+ and 75+ groups into one. We also provide the empirical CDFs of the distributions, and include both the weighted and unweighted distributions for the ACS.

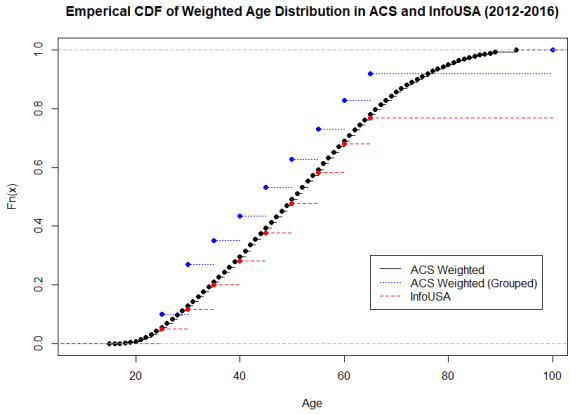


We can immediately see in this first set of distributions that the InfoUSA heads of household tend to be younger than those in PUMS. We also see that age in InfoUSA tracks closer to the weighted sample for PUMS than for the unweighted sample. Finally, note that from 2006-2010 to 2012-2016, the InfoUSA age distribution shifts up, with heads of household generally becoming older.


Emperical CDF of Age Distribution in ACS and InfoUSA (2006-2010)
Fn(x)
Age
ACS
ACS (Grouped)
InfoUSA


Emperical CDF of Age Distribution in ACS and InfoUSA (2012-2016)
Fn(x)
Age
ACS
ACS (Grouped)
InfoUSA


Emperical CDF of Weighted Age Distribution in ACS and InfoUSA (2012-2016)
Fn(x)
Age
ACS Weighted
ACS Weighted (Grouped)
InfoUSA


Emperical CDF of Weighted Age Distribution in ACS and InfoUSA (2012-2016)
Fn(x)
Age
ACS Weighted
ACS Weighted (Grouped)
InfoUSA