Housing Imputation

Introduction

The Housing Choice Voucher (HCV), Section 8 Project-Based Rental Assistance (PBRA), and public housing programs are the largest federal rental assistance programs administered by The Department of Housing and Urban Development (HUD). Smaller programs that we also consider include Moderate Rehabilitation (project-based), Section 221 Below Market Interest Rate (BMIR) (project-based), Section 202 Supportive Housing for the Elderly Program (project-based), and Section 811 Supportive Housing for Persons with Disabilities (project-based). For simplification, we lump these smaller programs into the PBRA program. These programs are intended to subsidize monthly rent payments for low-income families through long or short-term subsidy contracts for either (depending on the program) the tenant or owner of privately owned housing units, or through publicly owned housing units with lowered rents. For privately owned housing, these contracts can be attached to the tenant (tenant-based housing choice vouchers), or to the private housing owner (project-based housing choice vouchers or project-based rental assistance) to compensate a portion of the housing unit's rent.

The Current Population Survey (CPS) provides HCV and PBRA micro-data in its March Supplement in its "HLORENT" variable. These data should include the total number of HCV and PBRA recipients as well as their overall weighted-sum benefits with the variables HLORENT and "FHOUSSUB" respectively. Also, the CPS provides public housing micro-data with its "HPUBLIC" variable indicating the number of public housing recipients and FHOUSSUB conveying the corresponding weighted-sum benefits. However, Trudi Renwick from the US Census Bureau used proprietary matching information regarding the CPS and HUD administrative totals to find that the CPS misreports these variables in the following ways[1].

First, FHOUSSUB is not originally reported by CPS respondents due to the question's inherent complexity. Many individuals do not know the actual difference between market rents, and what they pay for their rent (effectively the subsidy amount). Thus, the CPS imputes this variable using 36 average housing subsidy amounts in the American Housing Survey (AHS) of households that fall into three income groups (>$10,000, $6,000-$9,999, and <$6,000), three household bedroom categories (>2 bedrooms, =2, <2), and four US regional categories. This method loosely calculates the housing subsidies as it lacks much of the US housing market heterogeneity. Although FHOUSSUB is poorly imputed, better imputation of this variable is beyond the scope of our project; thus, we use the CPS imputed FHOUSSUB variable. However, in our source code and results, we provide the option to use the Supplemental Poverty Measure's (SPM) dataset variable "spmu_caphousesub" for imputations, which is the SPM's better-calculated version of FHOUSSUB. This SPM dataset, which can be found here (for the 2013 version), is created by the US Census Bureau, and uses improved methods for the

1 Mitchell, Renwick, *"ESTIMATING THE VALUE OF FEDERAL HOUSING ASSISTANCE FOR THE SUPPLEMENTAL POVERTY MEASURE"*, SEHSD Working Paper,2016, https://www.census.gov/library/working-papers/2016/demo/SEHSD-WP2016-01.html, U.S. Census Bureau.

imputation of FHOUSSUB. These improved techniques for the definition of the housing subsidy value may provide more accurate and reliable imputation results. One can consult the paper "*Estimating the Value of Federal Housing Assistance for the Supplemental Poverty Measure*" cited below, for these more effective methods for calculating the housing subsidies for CPS respondents.

Second, only around 70% of CPS respondents reporting housing assistance (either HPUBLIC or HLORENT) actually did receive any type of housing assistance. This means that around 30% of CPS housing assistance responses are false positives, indicating respondent confusion regarding the housing assistance questions. Additionally, HPUBLIC is over-reported as 50% of those who report receiving public housing instead received rent subsidies through other programs (HLORENT); consequently, HLORENT is underreported. Due to over-reporting of both the overall housing assistance and specifically HPUBLIC for those who actually do receive assistance, we make the following assumptions. We assume that a significant amount of individuals reported receiving housing assistance in 2014, but did not actually receive it; consequently, we only keep recipients of each state that are the most likely recipients of housing assistance until the state recipient total matches the state administrative totals. Next, we lump the HPUBLIC and HLORENT variables together into one variable called "HOUSING", in order to correct for the respondents' confusion between these two variables.

We also noted that seventy-percent of those who have a positive subsidy amount are actual recipients of housing assistance; whereas only forty-percent of those who have a zero subsidy amount (FHOUSSUB = 0) actually receive housing assistance. Thus, we only consider HOUSING for those who have a subsidy amount greater than zero.

For a small number of states we found that HOUSING is underreported in the CPS compared to the administrative totals from US Department of Housing and Urban Development (HUD). Thus, we augmented the total number of recipients and their dollar amount of benefits from CPS 2014 March Supplement to match the 2014 Housing state-by-state reports provided by HUD.

We do not consider state-administered housing programs and their corresponding totals in our imputation as we do not have a reliable source of data on the extent of local and state housing programs, and these programs only constitute a relatively small portion of total housing assistance.
In addition to matching Housing recipients number and total benefits, we tried to maintain the original CPS micro-data distribution by utilizing HUD current distribution and Housing eligibility rules. However, since CPS data is insufficient in many ways, a number of assumptions are made (as explained above) in order to augment the data reasonably. This report details those assumptions and explains our imputation procedure.

CPS micro-data and Housing targets

In the Housing section of CPS March Supplement, the CPS imputes the market value of respondents' monthly housing subsidy: the variable that describes this monthly benefit is labeled as "FHOUSSUB". In addition, our created variable "HOUSING" represents whether or not a respondent

received either public housing (HPUBLIC), or HCV or PBRA (HLORENT), and whose imputed FHOUSSUB value is greater than zero.

Targets for imputation

The targets for imputation come from HUD official data. The HCV, PBRA, and public housing target data (both recipient and benefit amounts) come from HUD's "*Picture of Subsidized Housing*". We use "summary of all HUD programs" for the administrative totals by state. This summary contains administrative data on HCV, PBRA, public housing and the smaller programs we lumped into PBRA. The "summary of all HUD programs" contains the target aggregate administrative recipient amounts for all these programs by state, and their corresponding outlays, which serve as targets for HOUSING, and FHOUSSUB respectively.

In Fiscal Year 2014, administrative data suggests that approximately 4.5 million families claimed roughly 36 billion dollars in federal housing assistance benefits. The annual average combined benefit for each recipient is approximately $8,000 with significant variation across states (Table 2). Thus, for our imputation we partition the CPS March Supplement data by state for better measurements.

Below we provide a summary of CPS federal housing assistance outlays, and the corresponding administrative target outlays for each state before our imputation. See Appendix 1 for total recipients by state. Notice that FHOUSSUB is underreported for each state.

State	CPS total benefits (annually)	Admin total benefits (annually)
Alabama	201385761	503802936
Alaska	42450534	82680768
Arizona	147170596	292870344
Arkansas	143318073	233230632
California	1694363480	4801343616
Colorado	212622388	442967040
Connecticut	235105907	717290436
Delaware	62705601	104411160
District of Columbia	97219954	405067848
Florida	654725814	1514802780
Georgia	354226017	988780320
Hawaii	112606174	194505480
Idaho	67627412	70637424
Illinois	391039191	1792713552
Indiana	380272999	484271136
Iowa	103633819	183406080
Kansas	150750559	173821464
Kentucky	280032233	443840112
Louisiana	198295563	606897792
Maine	134597440	187453224
Maryland	244007658	940633008
Massachusetts	994836393	1961550000
Michigan	313029792	852089436
Minnesota	259541790	538243176
Mississippi	170593558	330956700
Missouri	340579322	499610244
Montana	90968013	69038892
Nebraska	63253386	130548132

Nevada	170023527	193823280
New Hampshire	74439015	164043600
New Jersey	514318471	1530092244
New Mexico	78756818	129233016
New York	2711384223	5316183072
North Carolina	417745116	718269024
North Dakota	45583695	55178496
Ohio	324405347	1367740080
Oklahoma	140540083	285799320
Oregon	254187175	342735780
Pennsylvania	878072822	1494576144
Rhode Island	74779153	289645104
South Carolina	91837700	372796656
South Dakota	44452202	64643400
Tennessee	330742926	578459952
Texas	875880853	1735369584
Utah	99961302	122109120
Vermont	37274796	95663712
Virginia	216767075	735742800
Washington	236454515	676742880
West Virginia	110816962	179899440
Wisconsin	321559485	363414240
Wyoming	30961906	31151796

Imputation Procedure

We follow a similar two-step procedure as the SSI imputation for both augmentation and reduction. First, we add up the family weights for each family in the recipient pool to see if we reach the administrative level for each state. If we don't, then we augment by including the most likely recipients from the non-recipient pool, using probabilities derived from both a logistic regression and a Random Forest classifier model, until we match administrative totals for each state. Second, we obtain an adjust ratio for the benefit amount that allows us to match the administrative dollar benefit totals for each state. If, upon aggregating the family weights in the recipient pool, we exceed the administrative totals for a state, then we reduce the number of recipients by choosing only the most likely recipients from the housing assistance recipient pool, using the same probabilities as the augmentation.

Step I: Recipient Imputation w/ Probabilities From Two Models

Model I:

We propose a basic logistic regression model for analyzing the likelihood of being a recipient. Following HCV, PBRA, and public housing eligibility rules, we use the available corresponding CPS information to create independent variables for AGI being less than 50% of the median, AGI being less than 30% of the median, a property value indicator (those who don't own a home have a zero value for this), family size, a disability indicator, an elderly indicator, a citizen indicator, a food stamps indicator for both family and household, a medical payment indicator, a Medicaid indicator. These variable are all important factors when determining Housing eligibility, since they are part of

the eligibility rules, or they were important variables listed in the Random Forest's feature importances.[2]

To create our income indicator variables, we used state-by-state median income tables downloaded from the HUD exchange site.

The included relevant independent variables are justified by the following income rules:

According to HUD, families can only be eligible for housing vouchers if their AGI is less than 50% of median income for their county. Since we do not have comprehensive CPS data on the county level, we take the average of all county median incomes within each state, to create state median income tables. In addition, each Public Housing Agency (PHA) must ensure that 75 percent of its admissions in each PHA fiscal year are families whose incomes are at or below the extremely low income limit: 30 percent of the median county income.

In addition, family size also helps determine income limits within counties. Elderly status, and disability status also contribute to the likelihood of eligibility.

We include food stamps, out of pocket medical care, and Medicaid participation, because food stamps and Medicaid participation is a proxy for indicating that a household is seeking means-tested government assistance on some level and has qualified, and because medical costs can be deducted from considered income amounts.

Lastly, we include property values, because CPS respondents don't report any value if they are renting, so all recipients of housing assistance must have the same "not reported" indicator.

Below we give our proposed logistic regression model for predicting the likelihood of receiving housing choice vouchers.

$$\begin{aligned} housingindicator = \alpha + familysize * \beta_1 + under30inc * \beta_2 + under50inc * \beta_3 + disability * \beta_4 \\ + elderly * \beta_5 + citizenship * \beta_6 + ffoodst * \beta_7 + hfoodst * \beta_8 + FMOOP * \beta_9 \\ + medicaid * \beta_{10} + propval * \beta_{11} + \varepsilon \end{aligned}$$

We then run the model on the CPS dataset. After, we use the fitted coefficients to produce a vector of probabilities for housing assistance recipients. We then rank all recipients according to their fitted probability. For each state sub-group, we aggregate the recipient weights, and add extra recipients by likelihood until the weights reach administrative level. For states whose pre-imputation weights are bigger than administrative targets, we similarly rank all recipients according to their probabilities, and only keep recipients who have the highest likelihood until we reach administrative targets.

Model II:

We use a Random Forest Classifier (RFC) model to determine housing assistance likelihood. Random

2 The code and results for the feature importances can be found in Rf_probs.ipynb

Forests performed much better than the logistic regression model, with an accuracy/score of .96, compared to the logistic regression's pseudo r-squared of 0.4.

To train the RFC model we used all of the CPS variables except those that uniquely , or approximately identified a household (household weights, ID numbers etc.). To create feasible variables for the training, we converted all variables containing categorical strings into numerical categorical variables, and created proxy variables for many columns with missing data (Not in Universe, None, etc.).

After training the Random Forest on a training set (80% of the data), we computed the probability that each CPS respondent received subsidized rent. Then, we ranked the probabilities as we did above, and imputed recipients until the recipient state totals matched the administrative state totals [3].

Step II: Benefit imputation

For each imputed/augmented recipient, we assign the average benefit amount for the corresponding state. Reduced/removed recipients' outlays are not considered .We then calculate the new total outlays for each state, and compare these outlays with HUD administrative state outlays. We calculate the adjustment ratios for each state by dividing administrative outlays by the new outlays . Most adjustment ratios close to 1, but some are significantly larger. We use these adjustment ratios to augment or shrink each household's benefits to match the state administrative totals.

3 The code and score results for the Random Forest Classifier model can be found in C-TAM's github documentation

Appendix

Table 1: Annual HOUSING recipient numbers by state for CPS and administration

State	CPS total family recipients	Admin total family recipients
Alabama	71734	84474
Alaska	17216	7268
Arizona	45490	38987
Arkansas	58469	46166
California	599405	458844
Colorado	83006	57678
Connecticut	74618	74999
Delaware	27134	11838
District of Columbia	36459	29251
Florida	253690	185365
Georgia	136370	124846
Hawaii	40253	18070
Idaho	29489	11726
Illinois	170762	201338
Indiana	161979	81036
Iowa	52250	36740
Kansas	72821	31558
Kentucky	104154	75948
Louisiana	83972	83872
Maine	44249	25074
Maryland	115066	89278
Massachusetts	316476	181625
Michigan	154190	134739
Minnesota	129654	85273
Mississippi	67475	50605
Missouri	149824	83771
Montana	30564	11961
Nebraska	35563	25359
Nevada	50620	21365
New Hampshire	24345	20650
New Jersey	167849	154181
New Mexico	32052	21582
New York	864736	551014
North Carolina	141716	118292
North Dakota	23987	10644
Ohio	161468	211071
Oklahoma	58674	49930
Oregon	68435	50551
Pennsylvania	266005	201534
Rhode Island	26251	35548
South Carolina	43482	59287
South Dakota	20849	11971
Tennessee	120160	101271
Texas	352786	255502
Utah	35435	18171
Vermont	13743	11672
Virginia	83569	94326
Washington	97767	84172
West Virginia	33697	31495
Wisconsin	164555	69780
Wyoming	11984	5161

Table 2: Average housing outlays by state

State	Average Benefit Amount
Alabama	5964
Alaska	11376
Arizona	7512
Arkansas	5052
California	10464
Colorado	7680
Connecticut	9564
Delaware	8820
District of Columbia	13848
Florida	8172
Georgia	7920
Hawaii	10764
Idaho	6024
Illinois	8904
Indiana	5976
Iowa	4992
Kansas	5508
Kentucky	5844
Louisiana	7236
Maine	7476
Maryland	10536
Massachusetts	10800
Michigan	6324
Minnesota	6312
Mississippi	6540
Missouri	5964
Montana	5772
Nebraska	5148
Nevada	9072
New Hampshire	7944
New Jersey	9924
New Mexico	5988
New York	9648
North Carolina	6072
North Dakota	5184
Ohio	6480
Oklahoma	5724
Oregon	6780
Pennsylvania	7416
Rhode Island	8148
South Carolina	6288
South Dakota	5400
Tennessee	5712
Texas	6792
Utah	6720
Vermont	8196
Virginia	7800
Washington	8040
West Virginia	5712
Wisconsin	5208
Wyoming	6036

Table 3: Adjustment ratios of outlays by state

State	Imputed	Admin	adjust ratio

Alabama	22745217	41983578	1.8458
Alaska	1956586	6890064	3.5214
Arizona	11068454	24405862	2.2049
Arkansas	10075632	19435886	1.9289
California	112890668	400111968	3.5442
Colorado	13454570	36913920	2.7435
Connecticut	20095934	59774203	2.9744
Delaware	2900790	8700930	2.9995
District of Columbia	7062234	33755654	4.7797
Florida	45785922	126233565	2.757
Georgia	27483682	82398360	2.998
Hawaii	4922890	16208790	3.2925
Idaho	2301926	5886452	2.5571
Illinois	54946309	149392796	2.7188
Indiana	16068039	40355928	2.5115
Iowa	6928561	15283840	2.2059
Kansas	6130011	14485122	2.3629
Kentucky	18884895	36986676	1.9585
Louisiana	16524630	50574816	3.0605
Maine	7423501	15621102	2.1042
Maryland	16717747	78386084	4.6887
Massachusetts	48700593	163462500	3.3564
Michigan	23774270	71007453	2.9867
Minnesota	15760730	44853598	2.8459
Mississippi	11790327	27579725	2.3391
Missouri	17295913	41634187	2.4071
Montana	3427338	5753241	1.6786
Nebraska	4103264	10879011	2.6513
Nevada	5957606	16151940	2.7111
New Hampshire	5697558	13670300	2.3993
New Jersey	40513147	127507687	3.1473
New Mexico	4958142	10769418	2.172
New York	149854240	443015256	2.9563
North Carolina	29234314	59855752	2.0474
North Dakota	2027790	4598208	2.2675
Ohio	54410040	113978340	2.0948
Oklahoma	10764738	23816610	2.2124
Oregon	17358270	28561315	1.6454
Pennsylvania	59260364	124548012	2.1017
Rhode Island	12478844	24137092	1.9342
South Carolina	16248418	31066388	1.9119
South Dakota	2283555	5386950	2.359
Tennessee	25387408	48204996	1.8987
Texas	58785729	144614132	2.46
Utah	5277628	10175760	1.928
Vermont	2775794	7971976	2.8719
Virginia	25060262	61311900	2.4465
Washington	16278769	56395240	3.4643
West Virginia	8629143	14991620	1.7373
Wisconsin	11249150	30284520	2.6921
Wyoming	1247579	2595983	2.0808

Table 4: Administrative and post-augmentation CPS totals

State	post augment CPS total benefits (annual)	post impute CPS total recipients	Admin total benefits (annual)	Admin total recipients

Alabama	503802936	83821	503802936	84474
Alaska	82680768	8155	82680768	7268
Arizona	292870344	37237	292870344	38987
Arkansas	233230632	45117	233230632	46166
California	4801343616	459908	4801343616	458844
Colorado	442967040	59607	442967040	57678
Connecticut	717290436	75576	717290436	74999
Delaware	104411160	12302	104411160	11838
District of Columbia	405067848	30219	405067848	29251
Florida	1514802780	186056	1514802780	185365
Georgia	988780308	122510	988780320	124846
Hawaii	194505480	19417	194505480	18070
Idaho	70637424	12716	70637424	11726
Illinois	1792713552	200896	1792713552	201338
Indiana	484271136	80286	484271136	81036
Iowa	183406080	37044	183406080	36740
Kansas	173821464	30690	173821464	31558
Kentucky	443840112	75306	443840112	75948
Louisiana	606897792	83972	606897792	83872
Maine	187453224	25388	187453224	25074
Maryland	940632996	88462	940633008	89278
Massachusetts	1961550000	181282	1961550000	181625
Michigan	852089436	132451	852089436	134739
Minnesota	538243164	86115	538243176	85273
Mississippi	330956688	51343	330956700	50605
Missouri	499610244	82851	499610244	83771
Montana	69038880	12402	69038892	11961
Nebraska	130548132	24721	130548132	25359
Nevada	193823280	21399	193823280	21365
New Hampshire	164043600	21514	164043600	20650
New Jersey	1530092232	154389	1530092244	154181
New Mexico	129233016	21079	129233016	21582
New York	5316183060	551859	5316183072	551014
North Carolina	718269024	115976	718269024	118292
North Dakota	55178496	11463	55178496	10644
Ohio	1367740080	212165	1367740080	211071
Oklahoma	285799320	50511	285799320	49930
Oregon	342735768	50267	342735780	50551
Pennsylvania	1494576144	202550	1494576144	201534
Rhode Island	289645092	35520	289645104	35548
South Carolina	372796656	59885	372796656	59287
South Dakota	64643388	12484	64643400	11971
Tennessee	578459940	101414	578459952	101271
Texas	1735369584	256475	1735369584	255502
Utah	122109108	18926	122109120	18171
Vermont	95663712	11985	95663712	11672
Virginia	735742788	94333	735742800	94326
Washington	676742880	83901	676742880	84172
West Virginia	179899440	31698	179899440	31495
Wisconsin	363414228	68788	363414240	69780
Wyoming	31151796	6152	31151796	5161