Quality of Match for Statistical Matches for the US

American Time Use Survey 2019, the Survey of Consumer Finances 2019, and the Annual Social and Economic Supplement 2020

Fernando Rios-Avila

This paper describes the quality of the statistical matching between the March 2020 supplement to the Current Population Survey (CPS) and the 2019 American Time Use Survey (ATUS) and Survey of Consumer Finances (SCF), which are used as the basis for the 2019 Levy Institute Measure of Economic Well-Being (LIMEW) estimates for the United States. In the first part of the paper, the alignment of the datasets is examined. In the second, various aspects of the match quality are described. The results indicate that the matches are of high quality, with some indication of bias in specific cases.

1 Introduction

This paper describes the construction of the synthetic dataset created for use in the estimation of the Levy Institute Measure of Economic Well-Being (LIMEW) for the United States. The LIMEW was developed as an alternative to conventional income measures that provides a more comprehensive measure of economic well-being. Construction of the LIMEW requires a variety of information for households. In addition to the standard demographic and household income information, the estimation process also requires information about household members' time use and information on a household's wealth, assets, and debts. Unfortunately, no single dataset contains all required data for the estimation.

In order to produce LIMEW estimates, a synthetic dataset is created combining information from three datasets, applying a statistical matching process. For current elaboration of this dataset, the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) 2020 is used as the base dataset, as it contains good information regarding demographic, social, and economic characteristics, as well as income, work experience, noncash benefits, and migration status of persons 15 years old and over. Time use data comes from the American Time Use Survey (ATUS) 2019, which provides rich data regarding how people divide their time among life's activities, including time spent doing paid and unpaid activities,

inside and outside the household, for one person in the household. Wealth data come from the Survey of Consumers Finances (SCF) 2019, which collects detailed information on household finances, income, assets, and liabilities.

This paper is organized as follows. Section one describes the data. Section two assesses the alignment of the information between ASEC and ATUS at the individual level, and the ASEC and the SCF at the household level. Section three briefly describes the methodology and analyzes the matching quality of the statistical matching. Section four concludes.

```
graph set svg fontface Garamond
```

<IPython.core.display.HTML object>

2 Data Description

2.1 Annual Social Economics Supplement (ASEC)

The CPS is a monthly survey administered by the US Bureau of Labor Statistics. It is used to assess the activities of the population and provide statistics related to employment and unemployment in the current labor market. Each household in the CPS is interviewed for four consecutive months, not interviewed for eight, and interviewed again for four additional months. Although the main purpose of the survey is to collect information on the labor market situation, the survey also collects detailed information on demographic characteristics (age, sex, race, and marital status), educational attainment, and family structure.

In March of every year, the previously interviewed households answer additional questions, part of the ASEC supplement formerly known as the Annual Demographic File. In addition to the basic monthly information, this supplement provides additional data on work experience, income, noncash benefits, and migration.

The ASEC 2020 is used as the base dataset (recipient), as it contains rich information regarding demographics and economic status. Because the time use survey (described below) covers individuals 15 years of age and older, younger individuals are discarded from the ASEC sample. This leaves us with a total of 124,806 observations, representing 264,870,782 individuals when weighted. For the household-level analysis, only information regarding the householder is used, leaving 50 observations, representing 150 households when weighted.

2.2 American Time Use Survey (ATUS)

The ATUS, a survey sponsored by the Bureau of Labor Statistics and collected by the US Census Bureau, is the first continuous survey on time use in the United States available since

2003. Its main objective is to provide nationally representative estimates of peoples' allocation of time among different activities, collecting information on what they did, where they were, and with whom they were.

The ATUS is administered to a random sample of individuals selected from a set of eligible households that have completed their final month's interviews for the CPS. The ATUS covers all residents who are at least 15 years old and are part of the civilian, noninstitutionalized population in the United States.

The ATUS 2019, which contains a total of 9,421 observations, is used as the donor dataset to obtain information regarding time use, which will be transferred to the ASEC 2020. Since information regarding household income is incomplete, the information was imputed using a univariate imputation process and information from the ASEC 2019. The sample represents a total of 248,718,989 individuals.

2.3 Survey of Consumer Finances (SCF)

The SCF is normally a triennial cross-sectional survey, sponsored by the Board of Governors of the Federal Reserve System in cooperation with the US Department of the Treasury, which collects information on families' balance sheets, pensions, income, and demographic characteristics.4 The purpose of the survey is to provide detailed information on households' assets and liabilities that can be used for analyzing households' wealth and their use of financial services.

In order to provide reliable information on household wealth distribution, the SCF is based on a dual-frame sample design. On the one hand, a geographically based random sample of respondents is interviewed to obtain a sample that is broadly representative of the population as a whole. On the other hand, a supplemental sample is obtained to include a sample of wealthy families in order to provide accurate information on wealth distribution, as the distribution of nonhome assets and liabilities is highly concentrated. In order to deal with the missing data, most variables with missing values are imputed using a multiple imputation procedure from which five replicates (imputations) for each record are obtained.

The SCF 2019 is used as the donor dataset to obtain information regarding assets, debts, and net worth. For the SCF 2019, a total of 5,777 families/households were interviewed. In order to account for the multiple imputation information, the five replicates are combined and used for the matching procedure. This provides a sample of 28,885 observations, representing 128,642,431 households when weighted.

3 Methodology

In order to create synthetic datasets that combine data from the different sources into a single dataset, we employ a methododology known as statistical matching.

The basic Statistical matching setup consists of having access to two sources of data: survey A and survey B, which collect information from two independent samples of the same population. Survey A collects information Z, X, whereas survey B collects information Z, Y. Although both surveys collect common information Z (for example demographics), they each contain information on variables that are not observed jointly: X (Consumption) and Y (Time use). In this case, the goal of Statistical matching is to create synthetic data that will contain all information Y, X, Z, linking observations across the datasets based on how close they are based on observed characteristics. It is also possible to constrain matches based on the weighted population each survey represents.

In turn, this synthetic dataset should allow researchers to analyze otherwise unobservable relationships between, X and Y, or as in our case, income and time use (**D. Orazio et al, 2006; Rasseler, 2002**). Thus, inference on the relation between X and Y can only be done to the extent that Z explains most of the common variation between X and Y.

3.1 Matching Algorithm

As described in **Lewaa et al (2021)**, statistical matching could be considered as a non-parametric variation of the stochastic regression approach, where no specific distribution assumption is imposed, and the imputed values are drawn directly from the observed distribution in the donor file. In particular, we implement a variation of the rank-constrained statistical matching described in **Kum and Masterson (2010)**, which improves on the approach by using a weight splitting approach in combination with clustering analysis for a better automatic selection of strata groups in the approach.[^ This is in contrast with previous iterations of Statistical matching used for the estimation of the LIMEW, which was based on ex-ante ad-hoc stratification, last-match-unit approach].

The statistical matching procedure applied for the paper is a multi-step process that can be explained as follows:

Step 1: Data harmonization

The first step involves harmonizing all common variables that will be used in the matching process and survey balancing. This is a necessary step in all imputation methods because variables need to have consistent definitions before they can be utilized for imputation using regression models.

Furthermore, this step includes adjusting sample weights to ensure that the weighted population is the same across all surveys. The standard practice is to adjust the sample weights of the donor sample. Additionally, for technical reasons, the weights are adjusted to whole numbers. While it is customary to adjust weights to match the total population, it may also be advisable to adjust weights to align with subpopulations based on selected strata variables.

Lastly, it is recommended to verify if both the donor and recipient files truly represent the same population by comparing the means, variances, and proportions of key variables across both surveys. In instances where significant imbalances are observed, reweighted methods can be employed to improve the balance between the surveys. However, there is no definitive rule to determine when a discrepancy in the distribution constitutes a substantial imbalance.

Step 2. Strata and Cluster identification, and propensity score estimation

The second step involves identifying statistically similar records based on observed characteristics Z. This is accomplished through a combination of three methodologies:

- **Principal Component Analysis (PCA)**: PCA is utilized as a data reduction technique to decrease the dimensionality of Z to a few linear combinations. While there are numerous suggestions on determining the optimal number of components, we simply select the first few components that explain approximately 50% of the data's variation.

- **Cluster analysis**: Once the principal components are estimated, they are employed to identify clusters within the dataset using a k-means cluster iterative partition algorithm. A brief description of the algorithm can be found in Hastie et al. (2009).

  As this algorithm only discovers locally optimal clusters and their identification is influenced by random initial conditions, it has a tendency to generate suboptimal clusters. To mitigate this issue, we modify the algorithm by repeating the procedure a sufficient number of times and selecting the "optimal" cluster based on the largest Calinski-Harabasz pseudo-F index (Calinski and Harabasz, 1974). This ensures that the chosen cluster maximizes intra-cluster similarity while minimizing inter-cluster dissimilarity.

  This procedure generates various sets of clusters of different sizes. The clusters with the highest number of groups are prioritized in the statistical matching procedure since they represent the most similar records, while clusters with fewer groups are utilized in later stages of the matching process.

- **Propensity score matching:** To enhance the matching procedure, we estimate a propensity score using a logit model. The dependent variable in the model is a binary indicator that determines whether an observation belongs to the donor or recipient file, while the independent variables consist of all common variables Z (including interactions or transformations). In the scenario where both surveys can be considered random samples from the same population, the expected coefficients for all variables should be zero or statistically insignificant. However, due to sampling variability and variations in survey design, it is common to observe variation in the propensity score. The logit model and propensity score can be estimated for the entire pooled survey or using the primary strata variables.

Step 3. Matching and weight splitting.

Once the propensity score has been estimated and the clusters and strata have been defined, we proceed with our matching algorithm. We rank all observations within each group, called cells, using the propensity score. The most detailed cluster, which defines the largest number of groups, is used for this purpose.

Within each cell, the record with the lowest propensity score from the donor file is matched or linked to the record with the lowest propensity score in the recipient file. If both records have the same weight, they are considered as fully matched and removed from the donor or recipient pool. If the weights are different, the record (donor or recipient) with the lowest weight is removed from the pool, and the weight of the matched record is adjusted by subtracting the weight value of the excluded record. The adjusted weight is retained in the pool for a subsequent match.

This process of matching records and adjusting weights, if necessary, continues until there are no more donor or recipient records left in that cell. If there are unmatched records from the previous steps, the procedure is repeated using a less detailed cluster until all records from the donor and recipient files are matched. Once the matching is completed, we obtain a synthetic dataset where all records in the donor file are matched to potentially multiple records in the recipient files, and vice versa.

For the final synthetic dataset, we select the "best" matched records, which are those that were matched in the earlier stages (most detailed clusters) or are records matched with the largest split weight. In case of ties, the "best" match is randomly chosen.

Due to this step, some observations in the donor sample may not be used at all, while others may be used more frequently than their weight would suggest. However, if the sample sizes and weight structures across both files are similar, we can expect only minor discrepancies between the distribution of the imputed data in the donor and recipient datasets. Nevertheless, if the sample sizes differ significantly, it is advisable to use the largest file as the recipient file.

The statistical matching procedure described above aims to impute all missing values in the recipient file by transferring the observed distribution of the imputed values from the donor file. This transfer of information is achieved by linking records from the donor file to records in the recipient files that have similar characteristics or similar ranks, based on an index known as the propensity score, which summarizes all those characteristics. Its main advantage over other methodologies is that constrained statistical matching fully recovers the distribution of the donor sample in the imputed sample, except for minor discrepancies as explained earlier.

4 Data Aligment and Statistics

4.1 ATUS 2019 – ASEC 2020

In order to create the synthetic dataset and transfer the time use information from the donor to the recipient dataset we need to verify that the distribution of information across surveys is comparable. To do so, we compare the distribution of individuals across the following set of variables: Gender, Parental Status, labor force Status, Presence of a Spouse, and employment status of the spouse. Since both datasets were collected within one year of each other, one should expect them to be well aligned.

Table 1: Summary Statistics, Alignment across Main variables

	ASEC	ATUS	diff
N	264870782	262823885	-0.8
Sex			
Female	51.4	51.6	0.2
Male	48.6	48.4	-0.2
Parental Status			
No	76.1	75.6	-0.5
Yes	23.9	24.4	0.5
Labor Force Status			
Not Employed	42.0	38.4	-3.6
Employed	58.0	61.6	3.6
Spouse present			
No	51.0	50.6	-0.4
Yes	49.0	49.4	0.4
Spouse's labor force status			
Spouse not Employed	18.1	17.3	-0.8
Spouse Employed	30.6	32.1	1.6

As can be observed in Table 1, the distribution of the sample with respect to sex and parental status is almost identical for both the ASEC and ATUS, with 48.5 percent of the sample being male, and about 24.1 percent being parents. The labor force status shows a relatively larger imbalance. The ATUS indicates there is a 3.6 percentage point larger share of employed individuals in the sample compared to the corresponding statistic in the ASEC survey (58 percent). The distribution of individuals across marital status presents a less severe imbalance. The statistics show that the share of married individuals is larger (3.6 percentage points) in the ATUS compared to the ASEC. In terms of the spouse's labor force status, the differences in the distribution among married individuals are negligible.

Table 2: Summary Statistics, Alignment across Selected variables

	ASEC	ATUS	diff
Household income category			
<$15k	7.3	8.6	1.3
$15k-$35k	16.1	16.7	0.6
$35k-$50k	12.7	12.7	-0.0
$50k-$75k	18.6	19.3	0.7
>$75k	45.3	42.7	-2.6
Age category			
15 to 24	15.8	16.2	0.4
25 to 34	17.1	16.9	-0.2
35 to 44	15.6	15.6	-0.1
45 to 54	14.9	15.3	0.3
55 to 64	15.9	16.1	0.2
65 to 74	12.2	11.8	-0.4
75 and older	8.4	8.1	-0.3
Race			
White	62.3	64.8	2.5
Black	12.0	11.9	-0.1
Hispanic	17.0	16.9	-0.2
Other	8.7	6.4	-2.2
Educational attainment			
Less than high school	13.9	14.4	0.4
High school	26.6	27.5	0.9
Some college	16.7	15.5	-1.2
College/grad school	42.8	42.6	-0.1
Number of children under 18 in household			
0	63.6	64.3	0.7
1	15.6	15.5	-0.1
2	13.2	12.4	-0.8
3	5.2	5.3	0.2
4 or more	2.5	2.4	-0.1
Number of persons in household over 18			
1	16.6	19.2	2.6
2	52.2	55.5	3.4
3	18.6	16.4	-2.3
4	8.5	7.0	-1.6
5 or more	4.0	2.0	-2.1

Table 2 presents statistics on additional variables that characterize the observations in both the donor and recipient datasets. The distribution across household income categories shows some imbalance, with the ATUS showing a considerably lower proportion of households in the highest income category, suggesting some undersampling of high-income households. For other demographic characteristics, such as age, race, and educational attainment, the distribution of individuals in both surveys is close. The largest observed differences across these characteristics are seen in the categories of some college (-1.2 percentage points) and whites (62.3 percentage points), with other differences falling below 1 percentage points. Finally, in terms of household structure, the survey's distribution is close in terms of number of children in the household, with slightly larger discrepancies in terms of the number of adult persons in the household, where the ATUS indicates a smaller share of larger households.

As expected, although some differences in the distributions can be observed between both surveys, these differences are small and there are no systematic differences that might seriously affect the quality of the matching process.

4.2 SCF 2019 – ASEC 2020

Table 3: Summary Statistics, Alignment across Main variables

	ASEC	SCF	diff
N	128451203	128642431	0.1
Household income category			
<$20K	15.4	14.9	-0.5
$20k-$50k	25.7	30.7	5.0
$50k-$75k	15.6	17.1	1.6
$75K-$100k	11.6	11.3	-0.3
>$100k	31.7	25.9	-5.8
Homeownership			
Renter	33.5	35.1	1.6
Owner w/Mortgage	36.6	42.1	5.5
Family type			
Owner wo/Mortgage	29.9	22.8	-7.0
Couple	55.1	56.0	0.9
Single Female	26.8	26.1	-0.7
Race category			
Single Male	18.1	17.8	-0.2
White	66.1	68.0	1.9
Black	12.8	15.7	2.9
Hispanic	13.7	10.9	-2.8
Age Category			

	ASEC	SCF	diff
Other	7.4	5.4	-2.0
<35 and younger	19.2	20.9	1.6
35 to 49	24.8	24.7	-0.2
50 to 65	28.0	27.9	-0.1
>65 and older	27.9	26.6	-1.4

Similar to the previous case, in order to create the synthetic dataset that combines the SCF and ASEC information, we use five strata variables to evaluate the aligment across surveys. These strata variables are income category, homeownership, family type, and race and age of the householder. In this case, the households/families rather than individuals are used as the unit of observation. Table 3 presents summary statistics that compare the distribution of observations within the strata variables. Since both datasets were collected within one year of each other, one should expect them to be well aligned.

As observed in Table 3, the distribution of households across income categories shows good balance across both samples, displaying at most a 5.8 percentage point difference when comparing households with highest income group. The SCF has a smaller share of middle-to-high-income households. This difference, however could be attributed to standard income growth across years. Nevertheless, for the matching, we aim to use income ranking rather than income levels to ensure good matching. Based on race and age, the distribution is very well balanced, with a less than 3 percentage point difference in the distributions, and a small underrepresentation of Hispanic and other races in the SCF. The largest distributional differences are present across family type and homeownership. The SCF dataset shows a slightly larger share of households within the "couples" categories (0.9 percentage points), while households with single females are underrepresented (-0.7 percentage points). Regarding homeownership, both samples present similar shares of renters and homeowners. Within the homeowners category, however, the ASEC underrepresents households with mortgages in about 5.5 percent of the instances compared to the SCF. To reduce the impact of the imbalance on the matching, we make use to include and indicator for "mortage" in the imputation process.

Table 4 presents statistics on additional variables that characterize the observations in both datasets. Information on education and occupation categories corresponds to that of the householder. The surveys are well balanced in terms of the educational attainment of the householder, the number of persons within the household, and the occupational categories. Although some imbalance is observed regarding the share of individuals with a Highschool degree.

Table 4: Summary Statistics, Alignment across Selected variables

	ASEC	SCF	diff
Education category			

	ASEC	SCF	diff
Less than high school	9.0	10.7	1.7
High school grad	27.2	24.5	-2.7
Some college	26.9	28.5	1.6
College or higher	36.9	36.3	-0.6
Sex of householder			
Male	72.7	73.0	0.3
Female	27.3	27.0	-0.3
Number of members in			
household			
1 Member	28.2	27.2	-1.0
2 Members	34.8	35.1	0.3
3 Members	15.1	16.1	1.0
4 or more Member	21.9	21.6	-0.3
Occupation category			
Occ1: 37–199	29.3	29.5	0.2
Occ2: 203–389	10.7	11.6	0.9
Occ3: 403–469 & 903–905	7.4	8.9	1.5
Occ4: 503–699	9.9	10.0	0.1
Occ5: 703–889	7.5	7.4	-0.1
Occ6: 473–499	0.7	1.2	0.6
Other	34.5	31.3	-3.3

5 Matching Quality

Statistical matching is a widely used technique, predominantly in observational studies in the medical literature. This method consists of combining the information from two separate and independent surveys into a single combined dataset from which statistical inferences can be obtained. The methodology enables the combination of the datasets using common information between both surveys, preserving the distributional characteristics of the combined information. In the following, the match quality between the ASEC 2020 (recipient) and ATUS 2019 (donor), and ASEC 2020 (recipient) and SCF 2019 (donor), correspondingly, are assessed.

5.1 Matching: ASEC 2020 and ATUS 2019

In order to obtain a good match, the matching process begins using five strata variables, namely sex, parental status, labor force status, marital status, and spouse's labor force status, to obtain 24 matching cells. Within each of these cells, propensity scores are estimated using logit models. A dummy variable indicating if the observation corresponds to the donor or the

recipient survey is used as a dependent variable. A set of demographic variables (i.e., age, educational attainment, race, parental status, marital status, and employment status) and household characteristics (i.e., number of adults, number of children, and household monthly income) are included as independent variables. For subsequent matching rounds, broader matching cells are defined accordingly, keeping the logit specifications consistent across all models, and including the omitted strata variable in the specification. The logit models and propensity scores are estimated using all information within broader cells, but the matching is done only across observations left unmatched from previous rounds.

Turning to the results of the match performance, Figure 1 shows the distribution of the matched records by matching round. As expected from these types of processes, 60.7 percent of the matches occur on the first round, ensuring the highest level of match quality. At the same time, 6.4 percent of the weighted sample was left unmatched after ten matching rounds. These unmatched observations should not bias the distributional statistics of the transferred information.
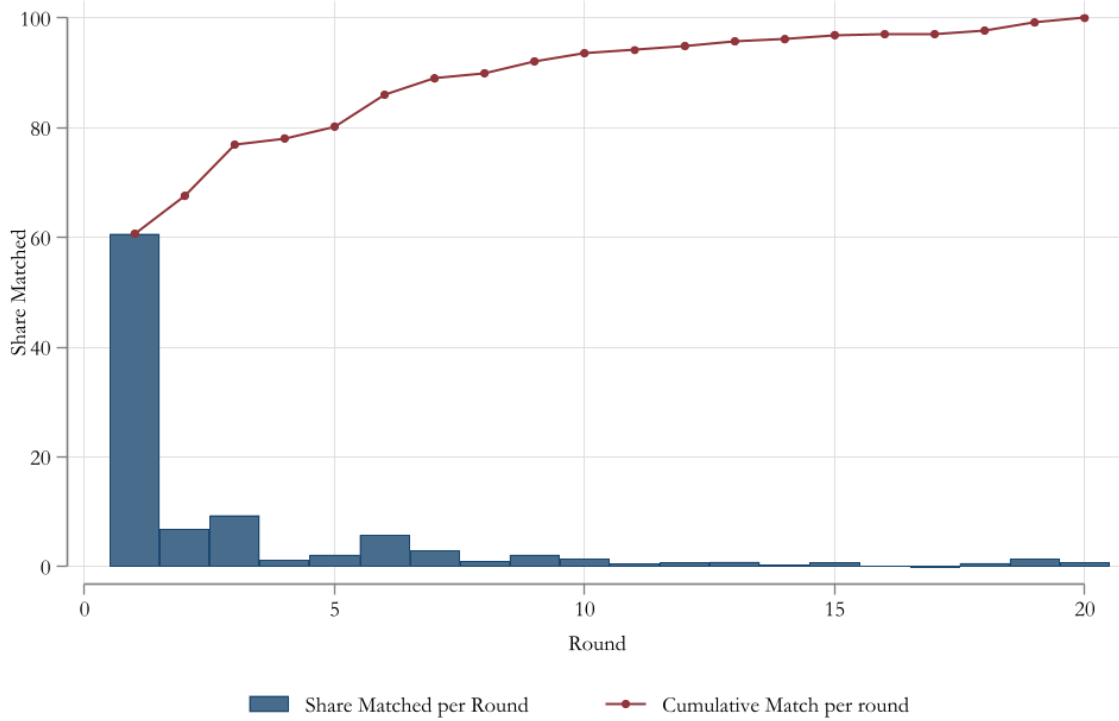
Figure 1: Distribution of Matched Records by Matching Round

Table 5 provides a description of the match quality, comparing some distributional statistics on the weekly hours of household production between the original information (ATUS) and the imputed data (ASEC). The table also compares the mean and median for the subcomponents:

Care, Core and Procurement. Both samples use the unadjusted survey weights. Given the large presence of zero hours allocated to household production in the sample, we provide the share of none-zero observations with hours of household production, constraining the sample when providing the relevant summary statistics. The percentile ratios are all equivalent with almost identical Gini coefficients (0.457). The means and medians on the disaggregated components of household production also show a strong equivalence between both surveys, indicating a strong balance in aggregate terms. This is expected by construction.

Table 5: Matching Quality: Summary Statistics

	ASEC	ATUS	Ratio ASEC/ATUS
Hours of Household Production			
Share with >0 Hrs	87.6	87.7	-0.10
p10	3.5	3.5	0.00
p25	7.8	7.6	3.08
p50	19.3	18.8	2.48
p75	36.5	36.2	0.97
p90	56.0	55.7	0.63
Gini	0.457	0.458	-0.06
Hour on Care			
Share with >0	23.2	23.7	-2.29
Average hrs per week if >0	14.5	14.4	1.16
Median hrs per week if >0	9.9	9.9	0.00
Hour on Core			
Share with >0	80.0	80.0	0.05
Average hrs per week if >0	16.4	16.2	1.20
Median hrs per week if >0	11.1	11.1	0.00
Hour on Procurement			
Share with >0	43.8	43.8	-0.01
Average hrs per week if >0	12.0	12.0	0.39
Median hrs per week if >0	8.8	8.8	0.00

Figure 2 presents a visual representation of the distribution of hours allocated to household production using three of the strata variables: sex, parental status, and labor force status. The figure shows that except for some values on the right tail of the distributions—for example, women who are not parents and are not working (F^P^W) and men who are parents and are not working (MP^W)—the overall distributions within the strata variables are analogous, indicating a good match quality.

For a detailed review of the performance of the matching, Figure 3 shows the ratios of the disaggregated hours allocated to household production (care, procurement, and core) between
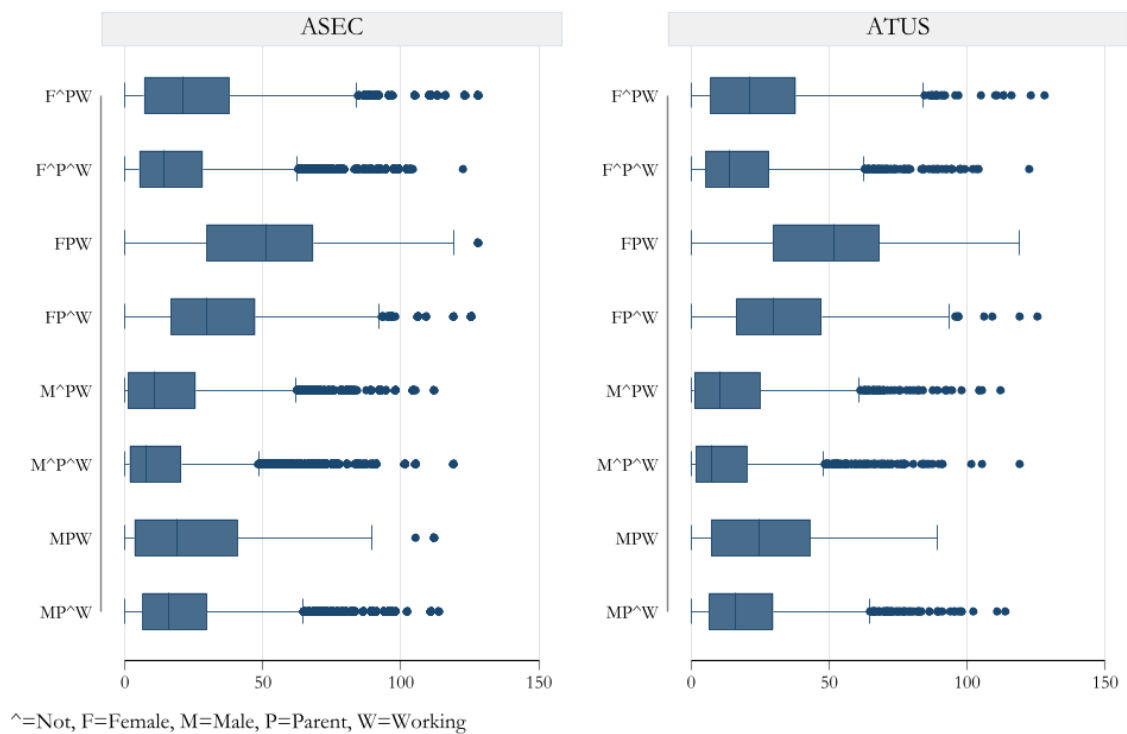
^=Not, F=Female, M=Male, P=Parent, W=Working

Figure 2: Distritubtion of Hours in Household Production, by Survey

the imputed data (ASEC) and the donor data (ATUS). Table 6 provides additional information on the mean and median hours of household production per week, and how large they are compared to the ATUS counterpart estimate.

The information is shown across the various selected varaibles used for matching. With some exceptions, the ratios of mean weekly hours of household production (and subcategories) fall within 5 percent of difference across all variables, an indication of good match quality. The largest differences are observed among youngest individuals and among people with less than a high school education. In both cases the statistics indicate 5 percent and 10 percent more hours respectively allocated to household production. In perspective, while such differences seem large, they may have a small effect on other analyses since the difference in average hours allocated to hours of household production among the specific is small (just about two hours per week).
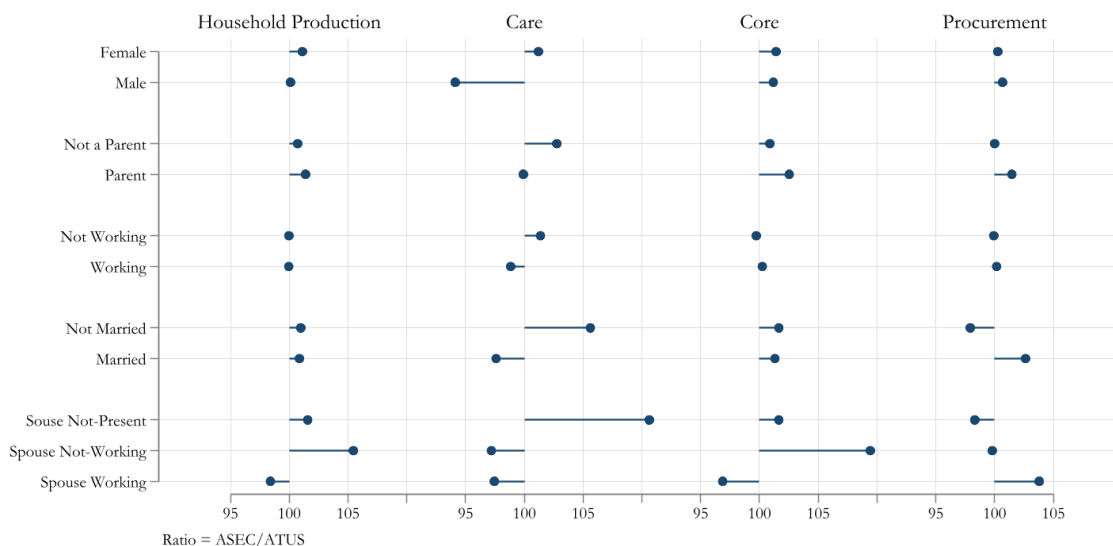
Figure 3: Ratio of Mean Household Production Hours, by main variables

Table 6: Average and Median Household Production Weekly Hours, by Selected Variables

	Mean	Ratio wrt ATUS	Median	Ratio wrt ATUS
Marital Status				
Not Married	17.5	101.0	10.7	98.9
Married	25.9	100.9	21.0	102.9
Parental Status				
Non-parent	18.9	100.7	12.8	100.0
Parent	30.8	101.4	25.2	101.4

15

	Mean	Ratio wrt ATUS	Median	Ratio wrt ATUS
Sex				
Female	26.6	101.1	21.0	100.0
Male	16.7	100.1	10.5	100.0
Labor Status				
Not Working	24.1	100.0	18.7	100.0
Working	20.0	100.0	14.0	100.0
Spouse's labor Status				
No Spouse present	17.7	101.6	11.1	100.0
SP not working	24.9	105.5	20.9	108.5
SP working	26.7	98.4	21.0	100.0
Education				
Less than high school	17.3	109.9	10.5	116.9
High school	21.4	98.0	14.0	92.3
Some college	21.1	101.9	14.6	108.7
College/grad school	23.7	99.6	18.3	98.1
Household income ($)				
<$15k	20.0	97.9	14.0	100.0
$15k-$35k	21.9	100.5	15.8	100.0
$35k-$50k	22.2	100.6	16.9	103.6
$50k-$75k	21.2	98.7	14.6	96.2
>$75k	22.1	101.8	15.8	103.1
Age Group				
15 to 24	12.7	104.9	7.0	109.1
25 to 34	21.7	97.1	14.6	96.2
35 to 44	25.5	97.7	19.7	99.4
45 to 54	23.0	106.5	17.5	115.4
55 to 64	22.3	102.1	16.4	100.7
65 to 74	24.5	95.1	20.2	96.1
75 and older	24.9	102.7	21.0	100.0

Figure 4 presents additional details on the quality of the match using the the combination of Sex, parent status, Labor status, and Spouse labor status. This figure plots the relative difference between CPS imputed hours and ATUS observed hours by group, againts the total hours observed in the ATUS. The size of the marker represents how representative a particular group is compared to the weighted sample. Larger markers indicating a larger share of the weighted sample.

Similar to the results described before, with some exceptions, total household production—in particular procurement and care hours—shows good levels of balance across most of the matching cells (note: procurement and core hours are part of household production). While
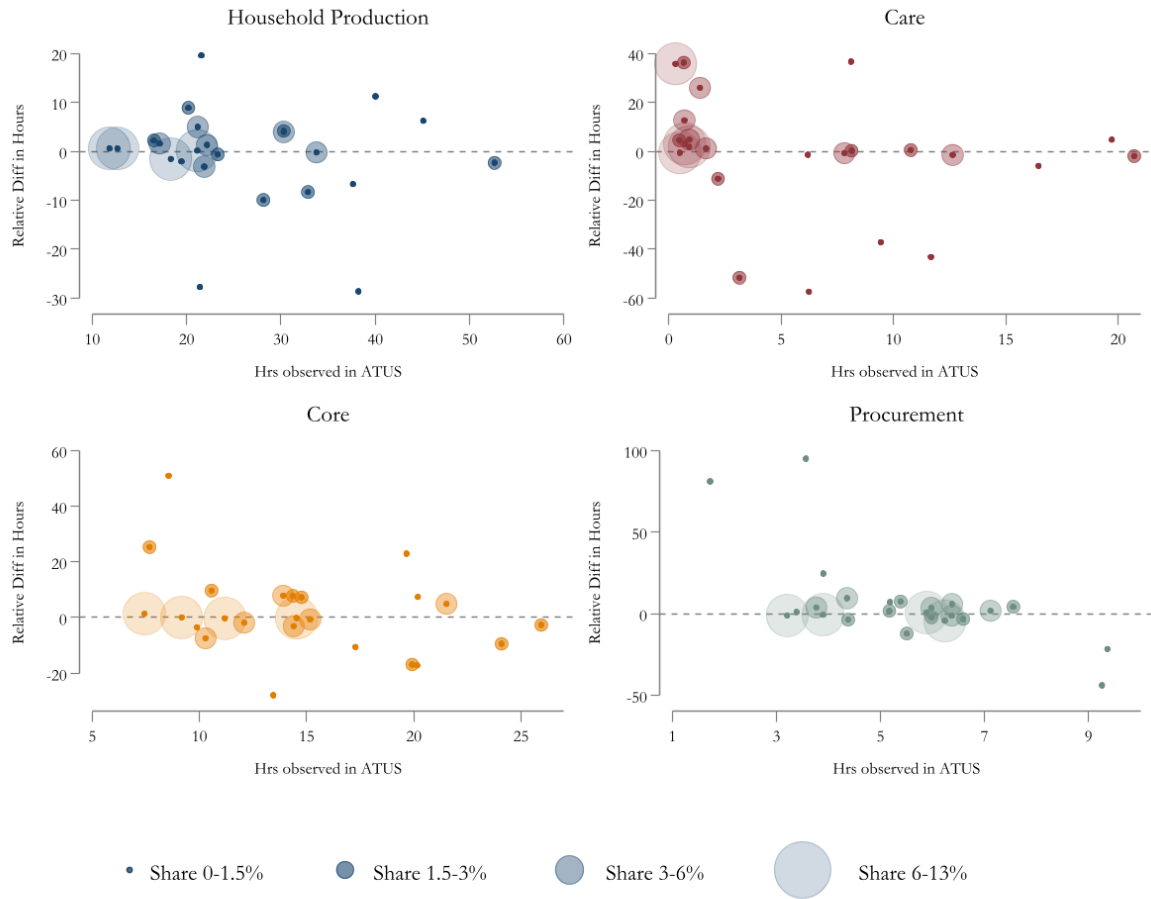
Figure 4: Difference of Mean Household Production hours by Cells

there are some differences larger than 20% of the ATUS data, those differences are observed to either small groups — as in the case in Procurement activities —, or for cases where observed hours of activities is low — as in the case of care activities.

To examine the match quality beyond the framework of the strata variables, Figure 5 presents information on ratios for household production and its components across education, household income level, and age group. The mean and median of total household production for these selected variables are provided in Table 6. In terms of education, people with high school and some college education have good levels of balance between both surveys. People with less than a high school education are imputed with longer hours allocated to household production (1.56 hours more) and all its components. In contrast, there is a small but consistent underestimation of the hours of household production (-0.09 hours) for people with at least a college degree. Individuals in the lowest income groups show an underestimation of the hours allocated to care activities (-0.27 hours).

In terms of age, the averages and medians indicate the statistical match did a good job imputing hours, as the differences are small for all groups. Looking at care activities, however, the statistical match seems to overestimate the number of hours spent on these activities, especially for people over 75 years of age and those younger than 24.
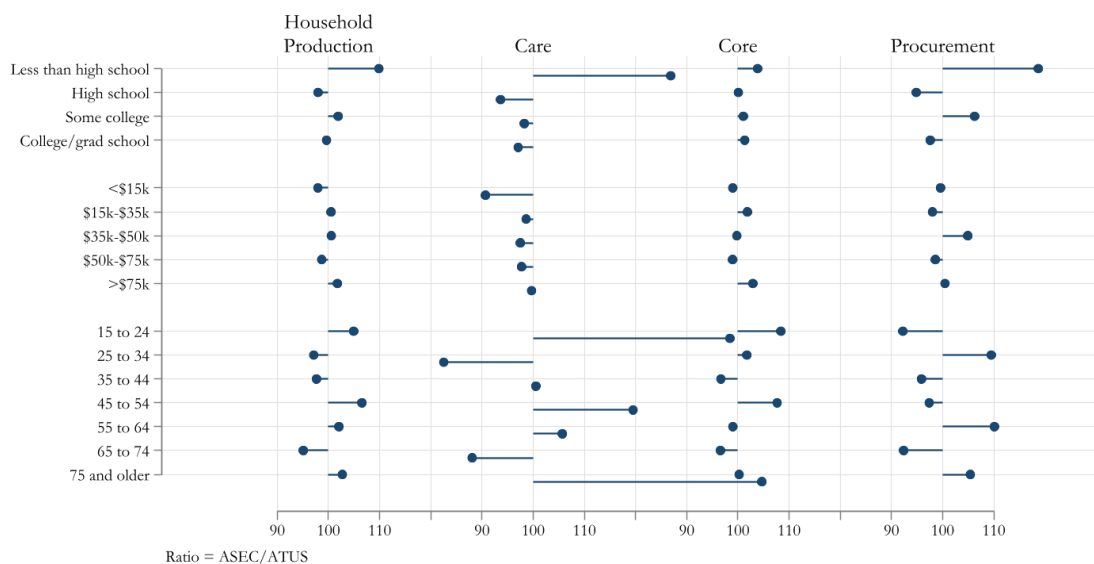

Figure 5: Ratio of Mean Household Production Hours by Selected Variables

5.2 Matching: ASEC 2020 and SCF 2019

For the matching process between the ASEC 2020 and SCF 2019, we consider 5 variables, namely income categories, homeownership, family type, and race and age of the householder

(head of household), as the most important to evaluate the quality of the matching.

For the estimation of the propensity score, in addition to the previously mentioned variables, a set of variables including dummies for zero income, zero wage income, dummies for other sources of income, age (and its square) of the householder, educational attainment, occupation category, and number of people in household are included in the model specification. They are used to create the propensity scores and principal components to elaborate the matching. Standardized indexes for income and wage income are also included.

Turning to the results of the match performance, Figure 6 shows the distribution of the matched records by matching round. As expected, a large share of the matches (51.3 percent) occurs on the first round, when the highest level of quality match is ensured. This is sufficiently large to obtain good matching quality in terms of the strata variables. Only 6.7 percent of the weighted sample is matched in the last round 3 rounds. This should not bias the distributional statistics of the transferred information in the aggregate.
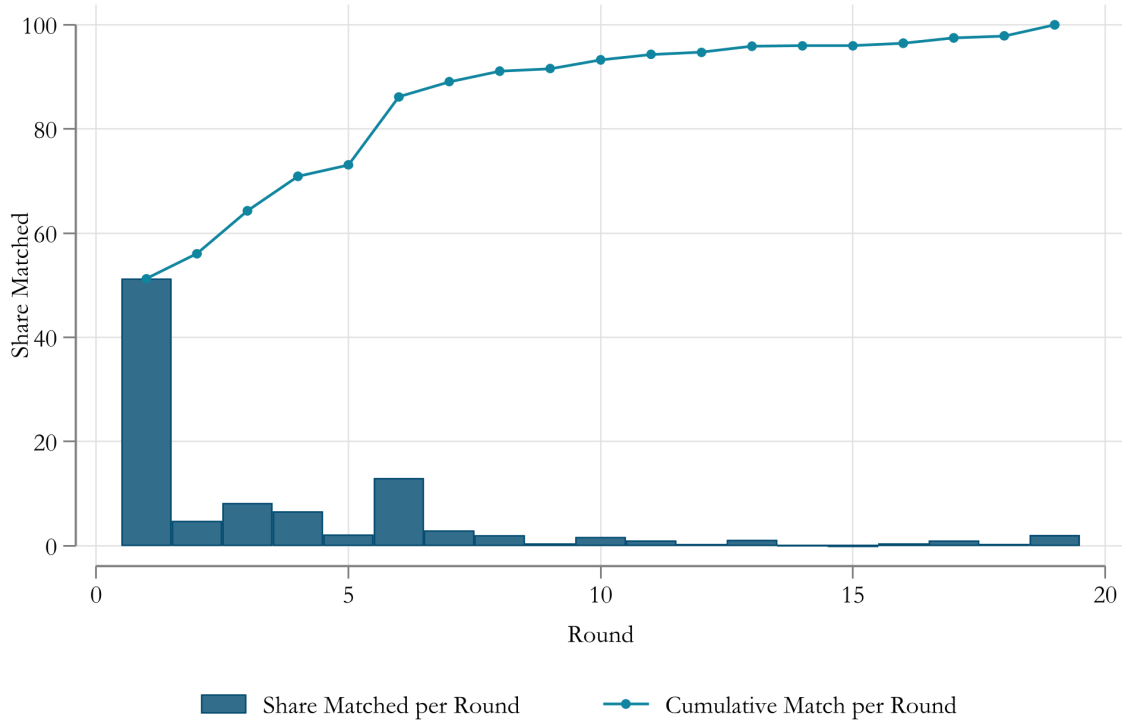

Figure 6: Distribution of Matched Records by Matching Round

Table 7 provides a better look at the match quality, comparing some distributional statistics on a household's assets and liabilities. Table 7 also presents some statistics on total assets and total debt. The upper percentiles and Gini coefficients are equivalent across both samples

(5.8 0.874). The lower percentiles, however, present a more pronounced difference, with the ASEC presenting lower net worth estimates. This is related to differences in the incidence of homeowners with mortgages shown in Table 3. The differences in the percentiles are also replicated when looking at the percentile ratios. The means and medians show a fair level of equivalence between both surveys for the disaggregated components. The largest difference corresponds to asset3 (liquid assets), showing an average difference of 5.8 4 percent, or about 5.8 $2,121.

Table 7: Matching Quality: Summary Statistics

	ASEC	SCF	Ratio ASEC/SCF
Distributional Statistics			
Networth			
p10	-10,761	-11,980	89.8
p25	2,000	1,481	135.0
p50	102,000	96,198	106.0
p75	377,000	367,300	102.6
p90	1,150,105	1,126,855	102.1
Gini	0.867	0.872	99.4
Summary Statistics			
Avg. Total Assets	802,317	804,556	99.7
Median Total Assets	205,000	198,505	103.3
Avg. Total Debt	97,781	97,435	100.4
Median Total Debt	25,000	24,000	104.2
Avg. Networth	704,536	707,121	99.6
Median Networth	102,000	96,198	106.0

Figure 7 presents a visual representation of the distribution of logged household net worth using three variables: race, homeownership, and age, using kernel density curves. Because there is a substantial share of the population who has zero or negative networth, we plot both sides of the distribution using a slight modification of the log transformation:

$$
\begin{aligned}
log\_networth &= log(networth) & & if\ networth > 0 \\
log\_networth &= 0 & & if\ networth = 0 \\
log\_networth &= -log(-networth) & & if\ networth < 0
\end{aligned}
$$

The figure shows that for most cases the distribution of the logged net worth is equivalent in both surveys. There are, however, some differences in the distributions regarding extreme values (outliers) among some groups, like households with white elderly homeowners (WHE), nonwhite elderly homeowners ($\widehat{W}$HE), or white nonelderly and nonhomeowners (W^{H}E). While

extreme values might not affect statistics like medians and percentiles, they might create problems when analyzing information at the means for more detailed subgroups. However, because the overall distribution is quite similar between donor and recipient sample across subgroups, the matching should suffice for less detailed analysis of the synthetic data.
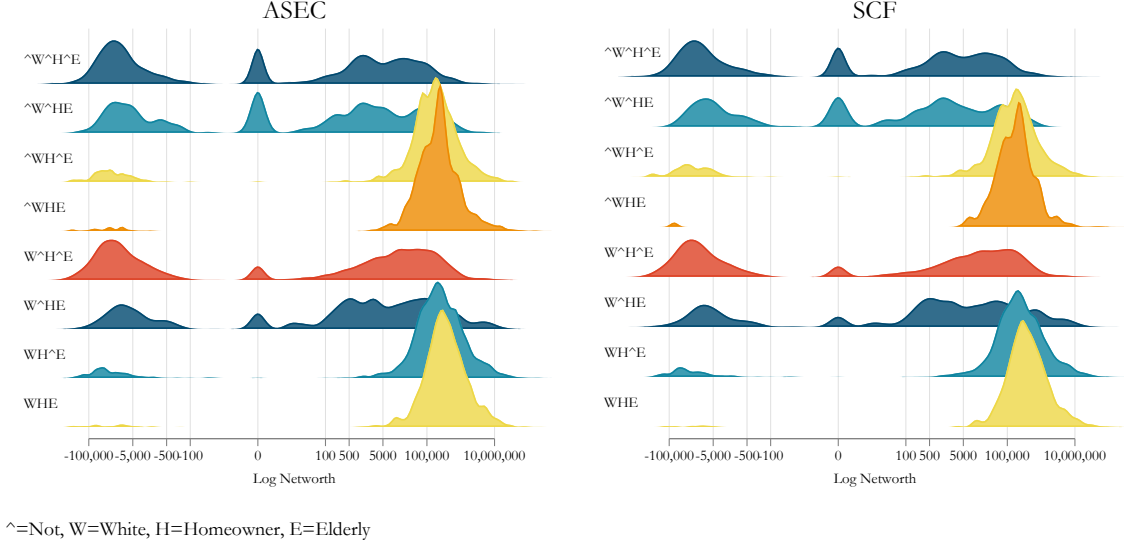

^=Not, W=White, H=Homeowner, E=Elderly

Figure 7: Distribution of Logged Net Worth, by Survey

For a detailed review of the performance of the matching, figures 6 and 7 show the ratios of asset and debt values between the imputed data (ASEC) and the donor data (SCF) across five variables used for the matching. Table 8 also presents information on the mean and median gaps of the net worth of the households with respect to the strata characteristics.

Table 8: Average and Median Networth, by Selected Variables

	Mean - ASEC	Ratio to SCF	Median - ASEC	Ratio to SCF
Total				
	704,536	99.6	102,000	106.0
Homeownership				
Renter	79,151	96.3	180	94.7
Owner with Mortgage	922,951	107.9	200,100	102.6
Owner w/o Mortgage	1,139,367	81.7	255,060	86.5
Income Group				
<$20k	166,951	232.7	16,290	1629.0
$20k-$50k	307,902	182.9	56,700	143.0
$50k-$75k	294,215	112.9	58,300	68.3

	Mean - ASEC	Ratio to SCF	Median - ASEC	Ratio to SCF
$75k-$100k	413,844	117.6	98,985	66.0
>$100k	1,593,584	73.7	288,400	56.5
Age				
Non-Eldery	582,776	101.0	66,650	111.5
Elder	1,002,769	95.4	220,301	94.5
Family type				
Couple	1,039,985	99.8	172,400	103.2
Single female	277,320	109.7	39,022	118.2
Single male	315,472	99.0	50,790	120.4
Race				
White	868,515	96.1	156,350	102.2
Black	302,369	250.8	11,200	138.3
Hispanic	262,167	152.4	21,500	143.3
Other	752,871	74.5	99,400	61.3

The first variable to be analyzed corresponds to the household income. After the matching, the average values of asset1, asset4, asset5, and net worth are overstated (up to 36 percent) in the recipient dataset among households in the lowest income group. This implies a difference of a little more than $7,411 for asset1 or 5.8 $11,140 for net worth. In contrast, with a few exceptions, all other assets/debts are understated in the imputed dataset by almost 5.8 10 percent on average, with the richest households having the largest bias (5.8 14 percent or 5.8 $227,000 lower net worth). In all cases, debt1 and debt2 are understated for all income groups except the richest, with a bias of less than 15 percent.

With respect to homeownership, the results show a good balance, on average, with net worth differences ranging from 5.8 $2,500 to 5.8 $79,500. The groups with the largest imbalances correspond to: homeowners without a mortgage, for which mutual funds (asset4) are understated by almost 22 percent and other debt (debt2) is overstated by 16 percent; and homeowners with a mortgage, for which mortgage debt (debt1) is overstated by about 11 percent and mutual funds (asset4) are overstated by 18 percent. In terms of family type, while households with couples and single women have well-balanced statistics, real estate assets in single-male households are understated by 35 percent (asset2) and mutual funds (asset4) are overstated by 13 percent. In aggregate, net worth is understated by 5.8 3 percent (5.8 $9,350) (Table 8).

When considering race, while the balance statistics show that information corresponding to households with white, black, and Hispanic householders is well balanced, the imputed sample consistently understates the asset/debt holdings from other race households by almost 5.8 17 percent. In terms of net worth alone, the net assets of "other races" are understated in just over 5.8 28 percent of the cases, which implies an almost 5.8 $151,731 difference. The median gaps show a much smaller absolute difference (5.8 $10,000), suggesting that the large
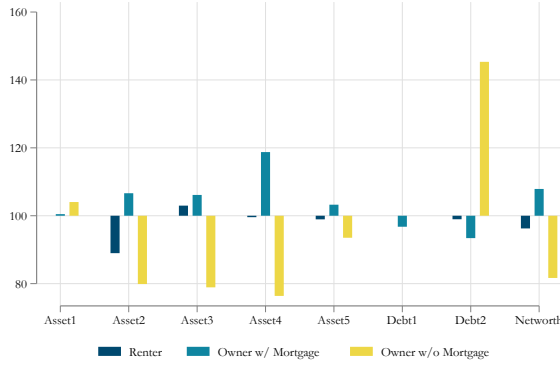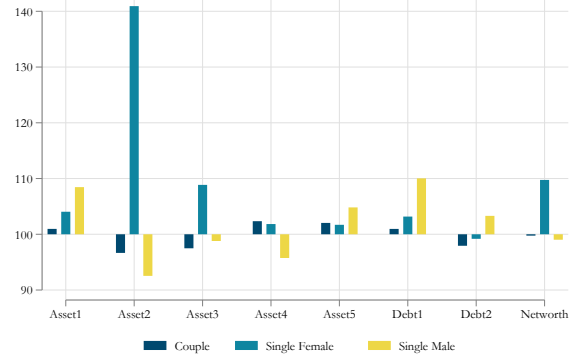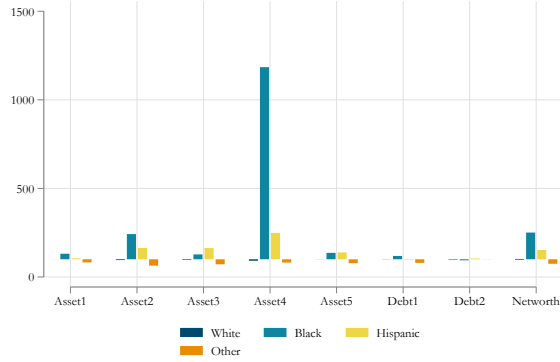
(a) Homeownership
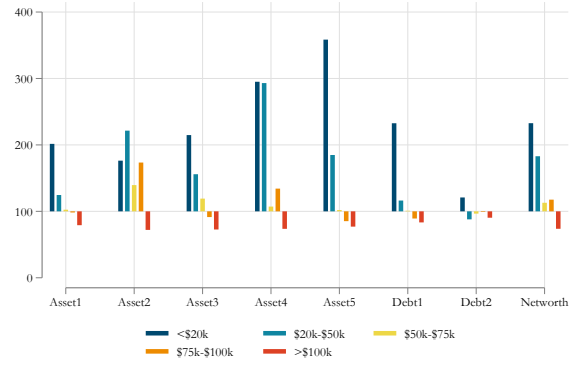
(b) Income group
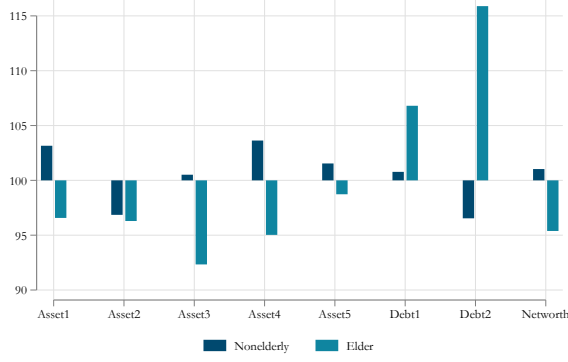
(c) Age

(d) Family type

(e) Race

Figure 8: Ratio of Mean Household Assets and Liability by Main Variables

differences in the means are driven by outliers. Finally, in terms of age groups, the statistics show that the imputed data is well balanced for most of the asset/debt components except for mortgage debt (debt1). The statistics show that the imputed data understates the debt of elderly households in about 11 percent of the cases. This happens because the share of elderly households with mortgage debt is lower in the ASEC survey compared to the corresponding share in the SCF.

To analyze how the matching performs for more detailed cells, the mean ratios between samples for all assets and debts are calculated for different combinations of the strata variables. @fig2p4 plots the densities corresponding to the mean ratios for selected combinations of the strata variables. As can be seen for most of the cases, the distributions of the mean ratios are highly concentrated around one, indicating that, on average, there is good balance between both surveys. As the figure also indicates, for some of the ratios, some large imbalances can be observed (ratios above two). These types of large imbalances for narrower cells are expected, as the SCF also collects information for high-income families, which might appear as large outliers. While for most variables the ratio distributions indicate a good balance, the ones corresponding to retirement assets (asset5) suggest that the imputed data tends to overstate it (25 percent).
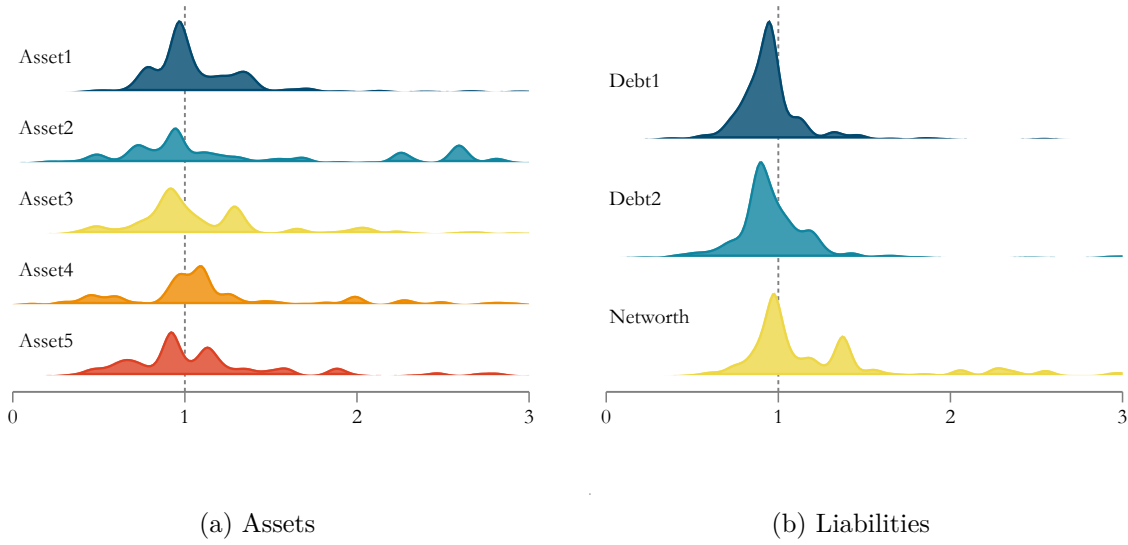

(a) Assets   (b) Liabilities

Figure 9: Kernel Density Ratios of Mean Household Assets and Liabilities

6 Conclusions

Overall, the ATUS and ASEC data are well aligned, with the some imbalances with respect to labor force status. The matching quality is good, with some limitations. There is a strong
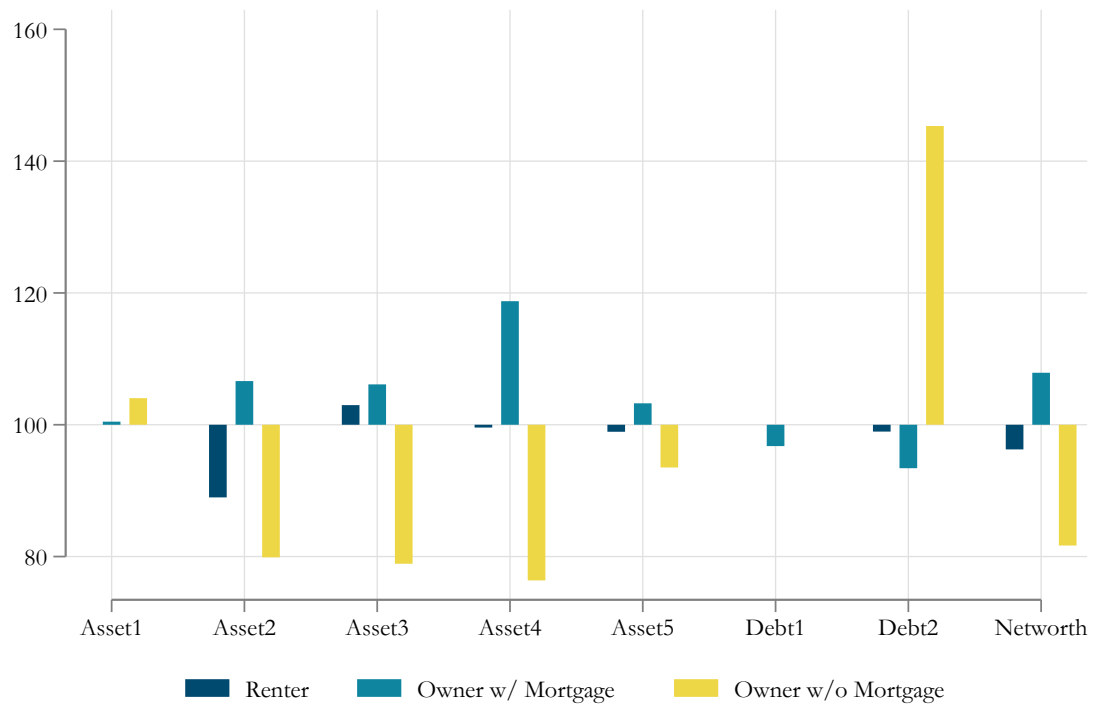
Figure 10: Homeownership

25

balance across the individual strata variables, showing good balance for aggregate measures (household production) for most of the variables analyzed. The results across the individual matching cells and other variables, however, show less balance.

On the one hand, the imputed information on the hours allocated to care activities shows important (relative) imbalances across many matching cells. The absolute differences, however, are small and should not create a large bias. On the other hand, information across other variables, such as education, household income, and (particularly) age, show important balance problems. The imputed dataset overstates household production of people with less than a high school education, and understates it for those with tertiary education, as well as for people in poor households. Across age, while the aggregate results are balanced, the individual components show large over- and under-estimations for different age groups.

With respect to the SCF and ASEC, the data is also well aligned, with the exception of mortgage holding, with small differences in the proportions of the breakdown by the sex of the householder. The results regarding the quality of the match are mixed. While the overall results show good balance between the imputed and donor surveys, with small underestimations of some items, analyzing the results across the strata variables shows relatively large imbalances (up to 5.8 20 percent) for a relatively small subset of strata variables. As we would expect, larger imbalances are observed for narrower groupings. The data shows some underestimation of mortgage debt, probably caused by the differences in the alignment of household property (see Table 3). Given that the SCF collects information from high-income households, it is possible that the information transferred from these observations has a strong influence on the cellspecific statistics. These results imply that careful consideration must be taken when making statistical inferences from certain populations. One can make inferences for the aggregate population, but attempting a similar analysis using two or more variables at the same time may carry too much bias to be informative.

7 References

Kennickell, A. B. 2000. "Revisions to the Variance Estimation Procedure for the SCF." Washington, DC: Board of Governors of the Federal Reserve System.

Kennickell, A. B., and R. L. Woodburn. 1999. "Consistent Weight Design for the 1989, 1992, and 1995 SCFs, and the Distribution of Wealth." Review of Income and Wealth 45(2): 193–215.

Kum, H., and T. N. Masterson. 2010. "Statistical matching using propensity scores: Theory and application to the analysis of the distribution of income and wealth." Journal of Economic and Social Measurement 35(3): 177–96.

Semega, J., and E. Welniak, Jr. 2013. "Evaluating the 2013 CPS ASEC Income Redesign Content Test." US Census Bureau Income Statistics Working Paper. Washington, DC: US Census Bureau.

Wolff, E. N., and A. Zacharias. 2003. "The Levy Institute Measure of Economic Well-Being." Levy Institute Working Paper 372. Annandale-on-Hudson, NY: Levy Economics Institute of Bard College.