Rios-Avila, Fernando

Working Paper

Quality of match for statistical matches using the Consumer Expenditure Survey 2011 and Annual Social Economic Supplement 2011

Working Paper, No. 830

Provided in Cooperation with:
Levy Economics Institute of Bard College

Suggested Citation: Rios-Avila, Fernando (2015) : Quality of match for statistical matches using the Consumer Expenditure Survey 2011 and Annual Social Economic Supplement 2011, Working Paper, No. 830, Levy Economics Institute of Bard College, Annandale-on-Hudson, NY

This Version is available at:
https://hdl.handle.net/10419/110006

Working Paper No. 830

Quality of Match for Statistical Matches Using the Consumer Expenditure Survey 2011 and Annual Social Economic Supplement 2011*

by

Fernando Rios-Avila†
Levy Economics Institute of Bard College

January 2015

* This work was supported by a grant from the Institute for New Economic Thinking.
† friosavi@levy.org

Levy Economics Institute
P.O. Box 5000
Annandale-on-Hudson, NY 12504-5000
http://www.levyinstitute.org

ISSN 1547-366X

Abstract

This paper describes the quality of the statistical match between the Current Population Survey (CPS) March 2011 supplement and the Consumer Expenditure Survey (CEX) 2011, which are used for the integrated inequality assessment model for the United States. In the first part of this paper, the alignment of the datasets is examined. In the second, various aspects of the match quality are described. The results show appropriate balance across different characteristics, with some imbalances within narrow characteristics.

Keywords: Consumer Expenditure Survey (CEX) 2011; CPS–March Supplement; Statistical Matching; United States

JEL Classifications: C14, C40, D31

Introduction

This paper describes the construction of the synthetic dataset created for use in the integrated inequality assessment (IIA) model for the United States. The IIA was developed to construct an alternative to conventional computable general equilibrium models, providing a more comprehensive method for assessing the impact of public policy on the distribution of economic well-being.[1] Construction of the IIA requires a variety of information for households. In addition to the standard demographic and household income information, the estimation process also requires detailed information on household consumption to capture household preferences and expenditure patterns.

In order to produce such estimates, a synthetic dataset is created by applying a statistical matching process in order to combine information from the Annual Social and Economic Supplement (ASEC) of the CPS 2011 and the CEX 2011.[2] The ASEC 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 people 15 years old and over. On the other hand, the CEX data is used to collect detailed information on consumption and expenditure patterns that will be used to estimate aggregate demand elasticities for the IIA project.

The IIA model also requires wealth data for households, which is not well captured in either the ASEC or the CEX. A statistical match of the ASEC with the Survey of Consumer Finances for 2010 was completed for the Levy Institute Measure of Economic Well-Being (LIMEW) project (Rios-Avila 2014), so we use the same match for construction of the IIA synthetic data set.

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

[1] For details on the IIA see Masterson et al., 2015 (yet to be published).
[2] For further details on the methodology see Kum and Masterson (2010). This paper uses a variation of the Statistical Matching process utilized in the construction of past statistical matching process (See Rios-Avila, 2014 and Masterson, 2014), which utilizes a split weight method (Kovacevic & Liu, 1994).

1. DATA DESCRIPTION

1.1. 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 such as employment and unemployment on 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, households interviewed answer additional questions, part of the ASEC supplement, formerly known as the Annual Demographic Supplement. In addition to the basic monthly information, this supplement provides detailed data on non-labor income sources and non-cash income sources, which are used to produce the official estimates on socioeconomic characteristics, such as poverty rates.

The ASEC 2011 is used as the base dataset (recipient), as it contains rich information regarding demographics and economic status. For the matching process, the household is used as the unit of analysis, and information regarding the householder and its spouse, as well as the household structure, is used for the matching process. This provides 75,148 observations, representing 118,682,616 households.

1.2. Consumer Expenditure Survey (CEX)

The CEX is a continuing quarterly survey of consumer units collected for the Bureau of Labor Statistics by the US Census Bureau. This survey consists of two surveys—the quarterly Interview Survey and the Diary Survey—which are used to provide information on buying habits, including data on expenditure, income characteristics. The Interview Survey is designed to obtain data on relatively large expenditures (i.e., property, durable goods, etc.). The Diary Survey is designed to obtain data on frequently purchased small items (i.e., food, housekeeping supplies, personal products, etc.), and is collected over a two-week period. This is the only federal survey that provides detailed information on a range of consumers' expenditure and income, as well as characteristics of consumers.

Consumption data for each consumer unit is collected quarterly over a 13-month period. The first interview is used for bounding purposes and not released publicly, while the remaining four quarters of data are used in this analysis. For the final sample, data from the raw FMLI files are used, as they contain demographic characteristics data on the consumer unit, which allows us to match the consumer expenditure survey to the CPS data. When all four quarters are aggregated, they provide a total of 26,990 consumption units, which represent 121,881,189 households.

2. DATA ALIGNMENT AND STATISTICS

2.1. ASEC 2011 – CEX 2011

In order to create the synthetic dataset and transfer consumption expenditure from the donor (CEX) to the recipient (ASEC) dataset, the first condition is for the datasets to approximately resemble the characteristics of the same population. For this, we start by aligning and comparing the definitions of key variables that characterize households. This is done to standardize the definition of different characteristics so they can be comparable across surveys, and compare if the surveys have similar marginal distributions, and thus come from the same population. Large misalignments between data would be worrisome as it would be an indication that the data are collected from somewhat different populations, and the matched data would not preserve the underlying relations (expenditure-demographics) observed in the donor data.

In Table 1, information regarding householder demographic characteristics, household structure, household income, and home ownership are compared between the ASEC and CEX surveys. The information corresponds to all the strata variables used in the matching process. While the purpose of this step is to standardize the definitions of key information across surveys, the ASEC and CEX present survey design differences that cannot be fully reconciled. In the CEX data, the final unit of analysis is the consumer unit, which is defined as a group of members of a household who are related or share at least two out of three major expenditures: housing, food, and other living expenses. In the ASEC, the unit of analysis is the household, which considers all members who occupy the same housing unit. As seen later, this causes some imbalances for specific variables, as more than one consumer unit could be present within a household unit. While there is no direct solution to this problem, the summary statistics (See Table 1) do not reveal any mayor consequences.

The second issue arises from the definition of "reference person." While both CEX and ASEC use the intrinsic concept of household ownership (or rent) as the principal criterion to identify the reference person, in the case of married couples, the ASEC randomly assigns one spouse as the reference person, whereas the CEX uses the owner of the household as the reference person. To reduce potential problems in the definition differences, information from both householder and spouse (if present) is used in the alignment and through the matching process.[3]

Overall, there seems to be an appropriate balance between both surveys, as they show similar distributions of households across surveys, with most variables showing less than a 1 percentage point difference, with some exceptions.

Table 1 Summary Statistics: Allignment Across Selected Variables

	ASEC	CEX	Diff		ASEC	CEX	Diff
Highest Householder Education				Men 16+			
Less than HS	9.8	11.2	-1.4	None	23.5	22.9	0.6
High school	25.9	22.3	3.6	1	64.3	63.8	0.5
Some College	28.6	31.6	-2.9	2+	12.3	13.3	-1.1
College	21.1	21.0	0.1	Women 16+			
Graduate School	14.5	13.9	0.6	None	15.7	16.2	-0.4
Oldest between Householder and spouse				1	70.4	68.4	1.0
16-29	12.0	13.4	-1.4	2+	13.9	15.5	-1.6
30-39	16.5	16.4	0.1	Men 2-15			
40-49	19.2	18.9	0.3	None	81.8	81.7	0.1
50-59	20.5	20.5	0.0	1+	18.2	18.3	-0.1
60-69	15.8	15.5	0.3	Women 2-15			
70+	16.0	15.4	0.6	None	82.5	82.6	-0.1
White Household				1+	17.5	17.4	0.1
Other	68.5	68.2	0.4	Children 0-1			
Both White	31.5	31.8	-0.4	None	94.1	93.7	0.4
Hispanic Household				1+	5.9	6.3	-0.4
At least one Hispanic	12.5	13.5	-1.0	Household Income			
Other	87.5	86.5	1.0	< 10K	7.8	8.7	-0.9
Family Type				10k-15k	6.0	6.7	-0.8
Husband & Wife only	21.6	20.8	0.9	15k-20k	6.1	6.4	-0.3
H/W with Children only	23.3	23.7	-0.4	20k-30k	11.5	11.8	-0.3
All other H/W households	4.0	4.8	-0.9	30k-40k	10.2	10.8	-0.6
Single parents	6.1	5.7	0.4	40k-50k	8.9	9.3	-0.4
Others CU	44.9	45.0	-0.1	50k-	14.5	13.8	0.8

[3] For simplicity, data of the most educated and oldest spouse are used in summary statistics.

				70k			
Family size				70k>	35.0	32.4	2.6
1 member	29.9	29.6	0.3	Household ownership			
2 members	32.2	31.4	0.8	Own	66.2	65.0	1.2
3 members	15.1	15.3	-0.2	Rent	33.8	35.0	-1.2
4+	22.8	23.7	-0.8				
Persons younger than 18							
No Younger	67.3	66.8	0.4				
1 person	14.2	14.0	0.2				
2+ persons	18.5	19.2	-0.7				
Persons Older than 64							
No older	75.4	75.5	0.0				
1+ older person	24.6	24.5	0.0				

Source: Author calculations based on ASEC 2011 and CEX 2011

Due to a combination of the reference person definition and household/consumer unit definition, there seems to be some misalignment regarding education attainment. Compared to the ASEC, the CEX suggests there is a larger share of householders with less than high school education and some college (a 2.9 and 3.6 pp difference). In addition, the ASEC shows a larger share of households with a household income larger than $70,000 (a 2.6 pp difference).

3. STATISTICAL MATCHING AND QUALITY

3.1. Methodology

Statistical matching (also known as data fusion) is a widely used technique in empirical studies, and has been applied in cases when no single survey contains all the relevant information needed for drawing important inferences. There are numerous empirical works that have applied this strategy in the economic field (See, for example, Rässler, 2002, and more recently, D'Orazio et al., 2006).

This method, which is similar to single imputation methods, consists of combining the information of two separate and independent surveys into a single combined dataset from which statistical inferences can be obtained. It enables the combination of the datasets using common information between both surveys, while trying to preserve the distributional characteristics of the combined information, under the assumption that both surveys represent the same population.

The algorithms that can be used to perform statistical matching can broadly be classified into two groups. The first one is known as unconstrained statistical matching. This strategy

frequently uses some type of distance criterion (propensity score matching, for example) so that the best possible candidate (based on observable characteristics) is chosen (often with replacement) from the donor file to be matched with the corresponding recipient observation. The second group is known as constrained statistical matching. In this case, the strategy imposes the restriction that all observations, specifically their weighted representation from both the donor and recipient surveys, need to be used in the final match. This strategy relies on a rank imputation, using broad strata variables to avoid undesirable matches.[4]

This paper uses a variation of the methodology proposed in Kum and Masterson (2010) (ranked CSM), which has been used in the estimation of the LIMEW.[5] The variation consists of using a weight-splitting strategy to better comply with the constrained statistical matching criteria.[6] The weight-splitting strategy implies that when two observations are matched, but their corresponding weights are different, the observation (donor or recipient) with the highest weight is "split" in two, to obtain one section that is perfectly matched with its counterpart, and the rest which is available for matching with the next observation in line. The advantage of this strategy is twofold. First, it allows the use of more information in the matching process, by allowing the creation of more detailed matching cells, which extends the matching quality control to a larger set of characteristics. Second, it perfectly complies with the CSM criteria, as all observations from the donor and recipient file are used (except for rounding errors).

3.2. Implementation

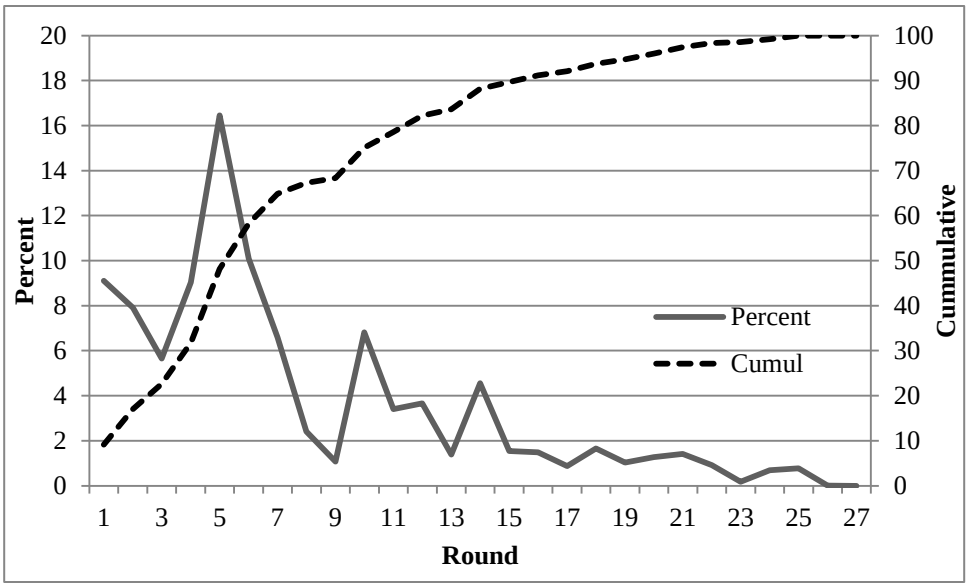
In order to obtain a good match, the matching process begins using 15 variables (all variables shown in Table 1), plus variables for region and whether or not the household is located in a metropolitan statistical area (MSA). This allows the initial creation of 5,062 matching cells with information available both in the donor and recipient data.[7] Propensity scores are estimated using a logit model for the whole sample. A dummy variable indicating if the observation corresponds to the donor or the recipient survey is used as a dependent variable, and the full set

[4] The hot deck matching uses ranked information based on some auxiliary information such as the propensity score. For further details on the matching procedure see Kum and Masterson (2010).
[5] For details on the LIMEW see Wolff and Zacharias (2003).
[6] In previous reports (see Rios Avila, 2013 and Masterson, 2010), the constrained statistical matching was partially achieved, as, in some cases, observations of the donor data were excluded from the analysis in early matching steps, leading to small unmatched data from both donor and recipient files. This suggestion is also made in Kum and Masterson (2010) for a better use of the full information from both donor and recipient data.
[7] A cell is defined as a group of observations (in this case, households) that have a set of common characteristics. A matching cell refers to a specific cell within which the matching process is performed.

of categorical variables presented in Table 1—variables for region, MSA indicators, age and age squared—are included in the specification as independent variables. For subsequent matching rounds, broader matching cells are defined, relaxing the exact matching criteria (i.e., using 2 instead of 4 race categories, or 5 instead of 8 income categories).

Turning to the results of the match performance, Figure 1 shows the distribution of the matched records by matching round. In 13 rounds, over 80% of the information was matched. Putting this into perspective, the 13th match is done using 14 education categories (husband and wife), 9 age groups, 2 race groups, 2 family types, 6 household income categories, 3 family size groups, and home ownership (owner or renter). This can be considered equivalent to the first round of the match process presented in Rios-Avila (2014) and Masterson (2014), with the advantage that observations matched previous to this point are matched with more detailed information. After 27 matching rounds, information from the donor file was fully matched to the corresponding recipient observation, and only 1 observation from the recipient file (that observation represents 37 households in the weighted data) was left unmatched.

Figure 1 Distribution of Matched Records by Matching Round: Weighted data



Source: Authors calculations based on ASEC 2011 and CEX 2011 data

3.3. Matching Quality

Since the constrained statistical matching is fully accomplished in this case, by construction of the modified matching process, all the moments of the expenditure data at the aggregate are perfectly aligned, and there is no added value in analyzing their distribution. Instead, the major

expenditure aggregates, namely Total Expenditure, Food, Housing, Transportation, Health care, Entertainment, Personal Care products and Services and Education, are summarized and compared across surveys using selected alignment variables (See Table 1).[8]

Table 2 presents information comparing the percentage differences of the expenditure aggregates with respect to selected variables used in the matching process. An overview of the results indicates that the estimates of total expenditure maintain a good alignment in the imputed survey, with a maximum difference of -6.4%. With respect to other aggregates, the results are also promising, showing less than a 10% difference with respect to the donor file. The largest differences are observed with respect to households with oldest and youngest couples, within single-parent households, and households with husband and wife present with members other than children.

[8] These aggregates are constructed and provided within the CEX data.

Table 2 Matching Quality Percentage Difference Between Imputed and Donor Expenditure: Selected Variables

	Total Expenditure	Food	Housing	Transportation	Health care	Entertainment	Personal care	Education
Education Householder Education								
Less Than HS	-2.5%	-4.2%	-0.8%	-2.6%	-8.0%	-0.2%	8.2%	72.2%
High School	-2.5%	-1.8%	-3.1%	-3.7%	-2.2%	-2.5%	-4.8%	45.2%
Some College	3.3%	2.7%	3.4%	4.4%	3.3%	3.5%	3.9%	-8.5%
College	-2.0%	-1.3%	-1.8%	-3.0%	-1.9%	-1.8%	-2.4%	-1.6%
Grad School	-0.4%	0.1%	-0.7%	0.8%	-0.1%	-0.7%	-1.2%	-2.7%
Age Householder and spouse								
15-29	6.0%	5.3%	6.3%	5.8%	11.7%	6.4%	8.3%	-11.7%
30-39	-2.5%	-2.9%	-2.9%	-1.7%	-3.3%	-3.0%	-3.0%	33.6%
40-49	-1.7%	-1.5%	-1.3%	-0.9%	-3.4%	-2.1%	-2.7%	-4.7%
50-59	-0.5%	0.4%	-0.4%	-0.7%	2.1%	-0.5%	-0.2%	0.3%
60-69	-0.7%	-0.9%	0.1%	-0.7%	-2.7%	-3.2%	-2.3%	15.4%
70+	1.9%	1.5%	0.8%	0.9%	-2.5%	7.2%	2.1%	25.2%
Family Type								
H/W couples only	0.6%	1.9%	-0.5%	4.1%	-1.1%	-2.0%	-1.4%	32.5%
H/W with children	-1.3%	-1.5%	-0.6%	-3.1%	0.1%	-0.5%	0.0%	-9.7%
H/W with other members	1.0%	-1.3%	1.1%	0.9%	-1.2%	5.7%	0.9%	-0.5%
Single parents w/Children	-5.1%	-5.1%	-5.4%	-4.3%	-3.3%	-1.7%	-11.0%	30.1%

All Other	2.3%	2.1%	2.2%	2.4%	1.9%	2.1%	3.0%	-1.8%
Family Size								
1	6.5%	7.6%	6.4%	7.7%	6.5%	6.0%	7.2%	-9.0%
2	0.0%	0.8%	-1.1%	3.4%	-1.7%	-1.9%	-2.5%	22.0%
3	-2.7%	-0.5%	-1.6%	-3.9%	-5.5%	-1.0%	-1.4%	-6.7%
4	-1.8%	-2.1%	-1.7%	-3.7%	2.3%	-4.3%	-4.5%	4.0%
5+	-0.8%	-5.7%	-0.3%	-3.5%	0.1%	6.7%	6.6%	-14.4%
Household Income								
< 10k	-0.5%	0.0%	0.4%	-2.1%	0.3%	1.6%	2.2%	-8.0%
10k-20k	0.3%	-1.7%	0.9%	-4.0%	0.8%	-2.6%	-2.1%	39.5%
20k-30k	-3.4%	-1.1%	-3.6%	0.1%	-7.4%	-5.2%	-1.1%	-13.4%
30k-40k	-3.4%	-2.1%	-2.2%	-5.5%	-0.4%	-3.6%	-8.6%	-7.3%
40k-50k	-4.0%	-2.9%	-4.1%	-3.7%	-6.4%	1.9%	2.5%	-0.5%
50k-70k	-3.8%	-2.3%	-3.3%	-3.3%	-3.2%	-5.3%	-4.2%	-0.9%
>70k	-3.1%	-2.4%	-2.7%	-3.2%	-1.4%	-3.5%	-3.8%	-4.0%
House Ownership								
Own	-1.3%	-0.9%	-1.1%	-1.7%	-1.7%	-1.4%	-1.6%	2.7%
Rent	1.9%	1.1%	1.7%	3.1%	2.9%	2.0%	3.0%	-6.3%

Note: The percentage differences is estimated as $\frac{E(Imp\ Exp_i)}{E(Actual\ Exp_i)} - 1$ for each group category i.

The only expenditure aggregate that shows large expenditure imbalances is education, with differences up to 160%. The imputed sample tends to overstate education expenditure among households with high school and less than high school education, middle (30–39 years) and old households (70+ years), and relatively poor households. In absolute terms, these groups present imbalances of $43 (less than high school) and $45 (70+ years old) per quarter.

While these large differences are initially worrisome, it is crucial to understand that education expenditure is a relatively special type of expenditure, as it is strongly related to the presence of school-aged children in the household. In this case, in order to determine whether the match/imputed sample is able to maintain the education expenditure structure with respect to the family structure, despite the observed imbalances, Table 3 presents similar statistics with respect to selected household structure information.

The information presented in Table 3 provides additional information on the quality of the match. Expenditures still show relatively low imbalances with some exceptions. For all expenditure categories, households with 3 or more women 16 years old and over tend to understate expenditures in the imputed data. This group, however, represents only 2.6% of the CPS data. Similarly, households with 3 or more men 16 years old and over tend to overstate expenditures on care, entertainment, and health care.

Perhaps more important is to observe that education expenditure shows a relatively better balance between imputed and actual expenditure, with fewer and smaller imbalances compared to those observed in Table 2. While the presence of these imbalances calls for caution when making inferences in specific groups, the results suggest that the final matched sample should provide appropriate information for inferences for most of the population.

Table 3 Matching Quality Percentage Difference Between Imputed and Donor Expenditure: Household Structure Variables

	Total Expenditure	Food	Housing	Transportation	Health care	Entertainment	Personal care	Education
Persons under 18yrs								
0	2.6%	2.7%	2.4%	3.7%	2.0%	2.2%	2.4%	0.2%
1	-3.6%	-0.6%	-3.3%	-4.2%	-5.9%	-1.3%	-5.9%	-4.3%
2	-2.8%	-3.3%	-3.1%	-2.7%	-2.9%	-3.5%	-2.9%	8.4%
3+	-5.5%	-8.8%	-2.8%	-9.9%	-5.2%	-4.9%	-2.2%	-7.3%
Persons 64+yrs								
None	0.0%	0.3%	0.2%	-0.2%	2.5%	-0.3%	0.0%	-0.2%
1+	0.0%	-1.0%	-0.6%	1.0%	-4.5%	1.2%	0.0%	3.1%
Men 16+								
None	1.5%	3.4%	0.1%	6.0%	-1.6%	2.2%	-9.9%	-8.8%
1	-0.4%	0.8%	-0.5%	0.3%	-0.2%	-1.3%	1.0%	4.2%
2	-0.5%	-4.2%	1.4%	-6.2%	-1.2%	0.6%	7.2%	0.3%
3+	2.0%	-11.8%	4.1%	-1.3%	20.3%	23.6%	19.6%	-25.9%
Women 16+								
None	4.6%	2.6%	5.0%	5.3%	10.9%	5.5%	29.3%	-8.6%
1	0.0%	0.7%	-0.5%	0.6%	-0.6%	-1.9%	-1.3%	15.1%
2	-2.6%	-3.0%	-1.5%	-2.7%	-5.2%	2.4%	-7.1%	-20.1%
3+	-8.0%	-10.8%	-6.9%	-16.7%	-7.1%	6.3%	-7.5%	-30.4%
Men 2-15								
None	0.8%	1.6%	0.5%	1.5%	0.3%	0.6%	0.9%	-1.2%
1	-2.0%	-3.0%	-1.0%	-4.2%	-0.6%	-2.2%	-5.1%	2.4%
2+	-5.1%	-10.4%	-3.4%	-7.6%	-3.2%	-1.7%	0.5%	18.1%
Women 2-15								
None	0.7%	1.0%	1.2%	0.1%	0.4%	1.5%	0.9%	-2.5%
1	-3.8%	-4.0%	-6.2%	-0.8%	-1.7%	-4.6%	-5.6%	17.5%
2+	-0.4%	-3.3%	-0.1%	0.4%	-2.6%	-6.9%	1.2%	-3.3%
Children 0-1								
None	0.0%	-0.5%	-0.1%	0.1%	1.0%	-0.9%	-0.1%	-1.9%
1+	-2.8%	3.2%	-3.7%	-5.6%	-9.0%	11.9%	3.7%	58.8%

Note: The percentage differences is estimated as $\frac{E(Imp\ Exp_i)}{E(Actual\ Exp_i)} - 1$ for each group category i.

4. APPLICATION: DETERMINANTS OF HOUSEHOLD EXPENDITURE

In the previous section, information was presented to show the quality of the matching process, showing the degree to which the matched sample is able preserve the marginal distribution of the expenditure with respect to different household characteristics. A drawback, however, is that those statistics provide little information on the joint distribution with respect to combinations of households characteristics.

In previous work (see Rios-Avila, 2014 and Masterson, 2014), such analysis was elaborated providing balancing information with respect to combinations of characteristics. While this strategy provides insights on the quality match for narrower characteristics, they tend to overstate localized imbalances (typically small subgroups or cases of skew distributions) and still ignore higher interactions and correlations of other variables that characterize the donor and recipient samples.

As an alternative, this paper opts to estimate simple linear models using both donor and imputed samples, where the dependent variable is the natural logarithm of total expenditure in the last quarter, and the independent variables are those used in the matching process. This should provide a better picture of the quality of the match, while taking into account multiple characteristics and providing some indication on the possible biases of inferences derived from the match data. Through the data, there are a few observations that report negative expenditures which are dropped from this analysis.

Table 4 Determinants of Total Expenditure: Log-Linear Model

	Donor			Imputed			t-test (p-val)
	Coef.	Std. Err.	t-stat	Coef.	Std. Err.	t-stat	
Household Education (base LTH)							
High School	0.1328	(0.0065)	20.35	0.1014	(0.0077)	13.24	2.36(0.0185)
Some College	0.1928	(0.0065)	29.51	0.2006	(0.0078)	25.79	0.58(0.5615)
College	0.3279	(0.0074)	44.08	0.3102	(0.0086)	35.97	1.17(0.2407)
Grad School	0.4472	(0.0083)	53.72	0.4323	(0.0092)	46.95	0.9(0.3673)
Age of Household (base 15-29yrs)							
30-39	0.0461	(0.0072)	6.41	0.0368	(0.0077)	4.80	0.66(0.5114)
40-49	0.0483	(0.0071)	6.77	0.0421	(0.0077)	5.43	0.44(0.6599)
50-59	0.0540	(0.0076)	7.08	0.0434	(0.0080)	5.46	0.71(0.4764)
60-69	0.0575	(0.0082)	7.01	0.0223	(0.0087)	2.58	2.2(0.0277)
70+	-0.0045	(0.0088)	0.51	-0.0192	(0.0092)	2.08	0.86(0.3887)
Family Type (Base H/W no							

Children)							
H/W with Children only	-0.0070	(0.0071)	0.98	-0.0263	(0.0075)	3.53	1.37(0.1697)
All Other H/W households	-0.0998	(0.0106)	9.46	-0.0669	(0.0116)	5.78	1.52(0.1282)
Single parents	0.0518	(0.0113)	4.59	-0.0134	(0.0120)	1.11	2.88(0.0039)
Others CU	-0.0525	(0.0075)	7.01	-0.0870	(0.0077)	11.25	2.33(0.0197)
Income category (Base <10k)							
10k-20k	0.2200	(0.0110)	20.09	0.1855	(0.0120)	15.43	1.6(0.1091)
20k-30k	0.4618	(0.0109)	42.36	0.4067	(0.0119)	34.29	2.57(0.0101)
30k-40k	0.6106	(0.0110)	55.27	0.5666	(0.0120)	47.16	2.03(0.0427)
40k-50k	0.6987	(0.0113)	62.04	0.6593	(0.0121)	54.56	1.79(0.0743)
50k-70k	0.8239	(0.0112)	73.86	0.7854	(0.0119)	65.73	1.77(0.0773)
>70k	1.1546	(0.0112)	102.84	1.1283	(0.0120)	93.85	1.2(0.2308)
Renter	-0.0743	(0.0045)	16.68	-0.0737	(0.0048)	15.44	0(0.9457)
Men 16+ (base: none)							
1	0.1516	(0.0071)	21.40	0.1182	(0.0073)	16.13	2.37(0.0178)
2	0.2310	(0.0086)	26.75	0.1836	(0.0089)	20.59	2.75(0.0059)
3	0.2698	(0.0121)	22.27	0.2130	(0.0134)	15.86	2.27(0.0233)
Women 16+(base: none)							
1	0.1413	(0.0077)	18.30	0.1121	(0.0080)	13.98	1.91(0.0558)
2	0.2204	(0.0088)	25.05	0.1786	(0.0092)	19.34	2.35(0.0187)
3	0.2615	(0.0125)	20.94	0.1833	(0.0127)	14.39	3.12(0.0018)
Men 2-15 (base: none)							
1	0.0700	(0.0054)	13.02	0.0573	(0.0059)	9.72	1.13(0.2599)
2	0.1437	(0.0076)	18.85	0.0822	(0.0080)	10.28	3.88(0.0001)
Women 2-15 (base: none)							
1	0.0637	(0.0055)	11.50	0.0580	(0.0058)	9.94	0.49(0.6261)
2	0.1120	(0.0080)	13.96	0.0915	(0.0086)	10.62	1.24(0.2165)
Any Children 0-1	0.0462	(0.0067)	6.92	0.0156	(0.0070)	2.23	2.23(0.0256)
Constant	7.9151	(0.0189)	417.82	8.0278	(0.0204)	392.79	3.82(0.0001)
R2	0.5630			0.5340			

Note: The model is estimated using a linear model. Models were estimated using sample weights. The t-test statistic corresponds to the difference between estimated parameters.

In Table 4, the parameters corresponding to the logarithmic expenditure model are presented, for both the donor and the imputed/matched data. In addition to the parameters, robust standard deviations and t-statistics are presented. It also includes a t-test for the null hypothesis that the estimated parameters for the donor and imputed data are equal to each other. The estimated parameters between the imputed and donor samples are similar for most of the

variables, which is corroborated by the low t-tests (p-values>5%), while the majority shows the differences not being statistically significant.

While the matched sample seems to offer good enough information for statistical inferences, as was previously shown in tables 2 and 3, there are some biases on the parameters which are statistically significant. On the one hand, there is an overall negative bias on the estimated marginal effect of household income. For instance, while households with $10,000 to $20,000 household incomes per year are estimated to spend about 0.22 log points more than the poorest households, in the imputed sample, the estimates indicate a lower increment of expenditure for the same income group (0.19 log points). While the estimates for each income category impact on total expenditure are larger and statistically different between the donor and imputed file, both show the same pattern on the estimates.

In a similar way, the parameters associated with the household structure (presence of men and women by age group) are consistently understated in the imputed sample, compared to the donor file, the largest statistical difference regarding households with 3 or more women (16+). It's worth mentioning that the constant (or base expenditure) is larger for the imputed data compared to the donor file.

5. CONCLUSION

This paper presents an application of a modified statistical matching algorithm that is used to combine data from the ASEC 2011 and the CEX 2011. Using a split-weight strategy allows us to introduce more information in the matching process, which helps improve the quality of the match to a larger segment of characteristics.

Overall, the ASEC and CEX data are well aligned, with some imbalances with respect to household education, family size, and presence of children (0-1) in the household. The matching quality is good, showing strong balance across different household characteristics, and showing good balance for total expenditure, as well as most expenditure aggregates, but with less accurate balance when analyzing education expenditure.

As an alternative to crosschecking expenditure balance for a combination of characteristics, a simple log linear model is estimated using the imputed (ASEC) data and the donor file (CEX) to assess the quality match and implication for statistical inferences when analyzing total expenditure. In general, the results are promising, with parameters that are

mostly statistically equal between each other. For some characteristics, like household income and structure, the results indicate a negative bias in the imputed sample.

While the estimation of the linear model is able to provide some information on the quality of the matched data for statistical inference, there are still some paths to explore. Given that statistical matching is in essence an imputation procedure, in its current state, the strategy used for this paper as well as the LIMEW indicators, relies on one set of imputed values for each recipient observation. Future research might need to explore the benefits for statistical inference of combining statistical matching with resampling techniques, such as bootstrap or multiple imputations.

References

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). Statistical Matching: Theory and Practice: Wiley.

Kovacevic, M. S., & Liu, T.P. (1994). Statistical Matching of Survey Data Files: A simulation Study. Proceedings of the Section on Survey Research Methods, 1, 479-484.

Kum, H., & Masterson, T. N. (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), 177-196.

Masterson, T.N. (2014). Quality of Statistical Match and Employment Simulations Used in the Estimation of the Levy Institute Measure of Time and Income Poverty (LIMTIP) for South Korea, 2009. Working Paper 793, Levy Economics Institute.

Masterson, T.N., Zacharias, A., Rios-Avila, F., and Kim, K. (2015) A Comprehensive Inequality Impact Assessment Methodology: An Application to Carbon Emissions Reduction Policies," Levy Economics Institute Working Paper No. forthcoming, Levy Economics Institute (yet to be published).

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. New York: Springer.

Rios-Avila, F. (2014). Quality of Match for Statistical Matches Using the American Time Use Survey 2010, the Survey of Consumer Finances 2010, and the Annual Social and Economic Supplement 2011. Working Paper 798. Levy Economics Institute.