

Using Administrative Data to Improve Nonresponse Weighting Procedures in the Consumer Expenditure Survey

Barry Steinberg (U.S. Bureau of Labor Statistics), John Voorheis (U.S. Bureau of
Census), Sally Reyes-Morales (U.S. Bureau of Labor Statistics) and Brett McBride (U.S.
Bureau of Labor Statistics)
2 Mass. Ave., NE, Room 3950, Washington, DC 20212
Steinberg.Barry@BLS.gov
John.L.Voorheis@Census.gov
Reyes-Morales.Sally@BLS.gov
McBride.Brett@BLS.gov

Abstract

The Census Bureau has begun research in collaboration with survey sponsors to improve the weighting procedures used in Title 13 household surveys by linking survey data to administrative data collected by other government agencies. The Consumer Expenditure Survey (CE) is one such survey where this research is ongoing, looking into the use of household-level income data collected by the Internal Revenue Service (IRS) to improve its nonresponse adjustment factors. Research shows that income is correlated with both response rates and expenditure reports.¹

Currently, CE uses publicly available zip code level data from the IRS to stratify its sample of households, but CE is looking into the use of non-public household-level data from the IRS to improve the stratification process. The IRS's publicly available data works fairly well. However, income varies within zip codes, and using IRS's non-public household-level data may improve CE's stratification process by putting the sample households into more homogeneous groups. This paper examines the impact of this linkage on CE's household weights and expenditure estimates.

Keywords: IRS, Nonresponse, Weighting, Data Linkage

1. Introduction

The Consumer Expenditure (CE) survey is a nationwide household survey conducted by the U.S. Bureau of Labor Statistics to find out how U.S. consumers spend their money. The target population is the U.S. civilian noninstitutional population. The CE Survey consists of two independent surveys: the CE Interview Survey and the CE Diary Survey. The CE Interview Survey collects detailed expenditure data on large or recurring expenditures such as refrigerators and utility bills, while the Diary Survey collects detailed expenditure data on small frequently purchased items such as food and apparel. Both surveys have the same sample design and data are collected by the U.S. Census Bureau.

The focus of this research is the CE Interview Survey, which is a nationwide rotating panel survey in which approximately 12,000 addresses are contacted each calendar quarter of the year. One-fourth of the addresses contacted each quarter are new to the survey. Of those

¹ John Sabelhaus, David Johnson, Stephen Ash, David Swanson, Thesia Garner, and Steve Henderson, *Is the Consumer Expenditure Survey Representative by Income?*(NBER Working Paper No. 19589, October 2013), pp. 1-2.

12,000 addresses, approximately 10,000 have occupied housing units, and usable interviews are obtained from approximately 6,000 of them. After an address has been in the sample for four consecutive quarters, it is dropped from the survey and a new address is selected to replace it.

CE data has many customers and the data are used in many ways. For example, the Consumer Price Index (CPI) uses the data to select the market basket of goods and services it uses to measure price inflation, and for its expenditure weights. The Internal Revenue Service (IRS) uses CE data to derive sales tax information for taxpayers who use the itemized deductions section of their tax returns. And the Supplemental Poverty Measure (SPM) uses CE data to determine poverty thresholds based on the amount of money consumers report spending on a basic set of goods that includes food, clothing, shelter, and utilities.

The CE surveys began using IRS's publicly available data in the weighting process in 2014 in the form of zip code level income. However, in 2016 CE started exploring the possibility of using IRS's non-public household-level income data to improve the stratification of the sample households in its nonresponse adjustment process. The IRS's publicly available data work fairly well in CE production with the higher income groups showing larger nonresponse adjustment factors², but since income varies within zip codes³, using its non-public household-level data may improve CE's stratification process by putting the sample households into more homogeneous groups. This paper examines the impact of using IRS's household-level income data instead of its zip code level income data on CE Interview Survey household weights and expenditure estimates.

2. Data Description

The research presented in this paper is based on CE Interview Survey data collected from April 2014 through December 2014. The dataset includes both respondents (households with completed interviews) and nonrespondents (households that could not be contacted or refused to give interviews). The dataset excludes residential addresses that are not occupied as well as nonresidential addresses. The dataset contains information from 36,638 records consisting of 19,425 respondents and 17,213 nonrespondents. This CE data was subsequently linked to administrative records from the IRS and to Census's Master Address File (MAFIDs) via a matching process and is described later in the paper.

3. Background

CE has a nonresponse adjustment process which adjusts the data to account for the nonrespondents. The process uses the traditional cell adjustment method where all the households in the sample are partitioned into 192 cells using variables based on a few demographic characteristics. The weights of the respondents in a cell are then increased to account for the nonrespondents by multiplying them by an adjustment factor equal to the inverse of the cell's response rate.

² The higher income groups using IRS AGI zip code data showed higher nonresponse adjustment factors when compared to the other income groups. Higher nonresponse adjustment factors are associated with lower response rates. The actual income groups are described later in the report.

³ Quentin Brummet, Denise Flanagan-Doyle, Joshua Mitchell, John Voorheis, Laura Erhard, and Brett McBride, CARRA Working Series Paper 2018-01, pp. 41-42

Nonresponse adjustment variables are a major component of the CE weighting process. Having good nonresponse variables can reduce the impact of nonresponse bias, which can be large when the respondents and nonrespondents have different characteristics. A successful nonresponse adjustment process depends on having variables of some kind that are available for both respondents and nonrespondents that are correlated with both propensity to respond and key survey outcome variables. This is an issue since in general relatively little information is available on households that do not respond to the survey.

Prior to 2014, the nonresponse adjustment variables in the CE Interview Survey were Region of Country, Consumer Unit Size, Housing Tenure (Owner/Renter), Race, and Rotation Group. However, Sabelhaus et al.⁴ examined CE's nonresponse adjustment process and concluded that:

- High-income households were under-represented and low-income households were over-represented in the CE Survey.
- High-income households under-reported both their incomes and expenditures.
- CE's weighting procedures should have made appropriate adjustments to account for the different response propensities but they did not.

Based on that paper, CE grew concerned that high-income households were under-represented and low-income households were over-represented even after adjusting their weights, and that this needed to be addressed. After analyzing the paper's findings, CE examined its nonresponse weighting variables and decided to add an income variable to its production weighting process, and to make a few other changes to its list of nonresponse variables:

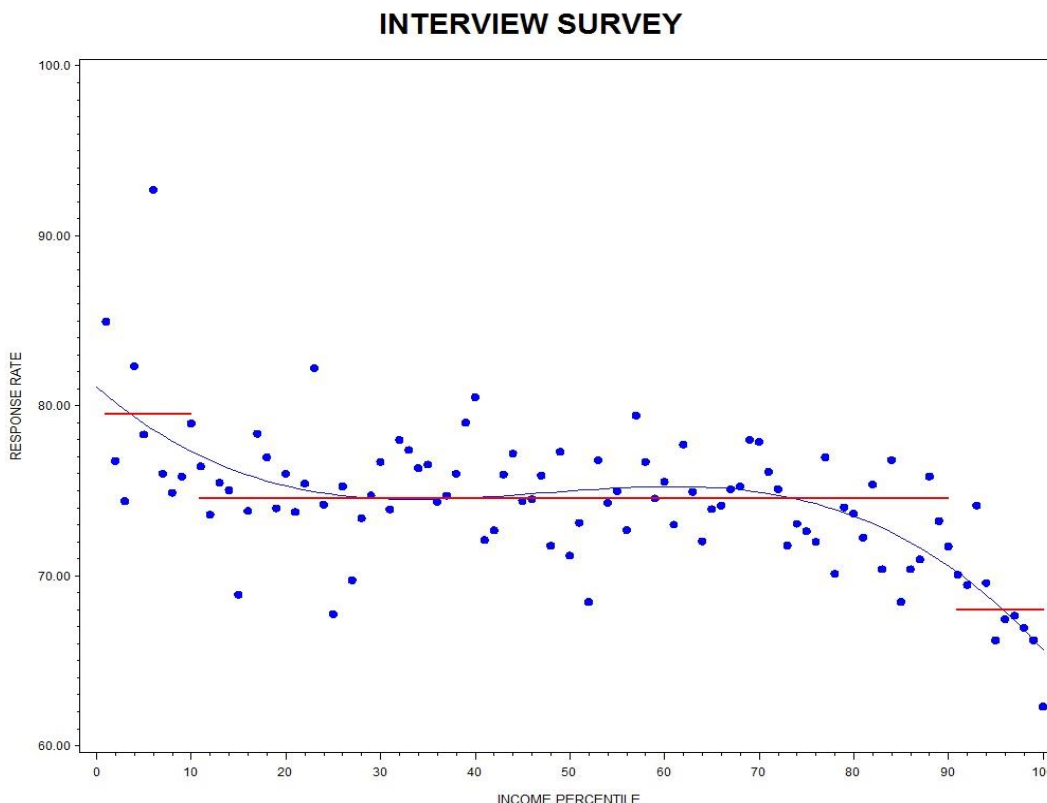
Nonresponse Adjustment Variables Used in the CE Interview Survey	
1986 – 2013	2014 – present
1. Region of Country	1. Region of Country
2. Consumer Unit Size ⁵	2. Consumer Unit Size
3. Housing Tenure *	3. Number of Contact Attempts **
4. Race *	4. IRS Adjusted Gross Income (AGI) by Zip Code **
5. Rotation Group *	
* discontinued in 2014	** new starting in 2014

⁴John Sabelhaus, David Johnson, Stephen Ash, David Swanson, Thesia Garner, and Steve Henderson, *Is the Consumer Expenditure Survey Representative by Income?* (NBER Working Paper No. 19589, October 2013).

⁵A consumer unit is a group of people living together in a housing unit who are related by blood, marriage, adoption, or some other legal arrangement; who are unrelated but pool their incomes to make joint expenditure decisions; or is a person living alone or sharing a housing unit with other people but who is financially independent of the other people. In most cases, consumer units and households are the same thing so the terms are often used interchangeably. For nonresponding addresses, the CU size variable is imputed by Census Region using the imputation proportions file. This file is created once a year prior to processing CE Interview Survey data for April. The 'source data' used to create it consists of one year's worth of interviews from respondents using processed CE Interview Survey data from April to March of the previous year.

When looking at the IRS adjusted gross income (AGI) variable, CE considered many factors but ultimately decided to classify it into three groups based on its relationship to response rates during the 2006-2010 time period as shown in Figure 1, below:

Figure 1. Response Rates for CE Interview Survey by Income Percentile: 2006-2010



Based on this graph, CE decided to classify IRS AGI by zip code into three distinct groups based on percentiles: the bottom 10%, the middle 80%, and the top 10%. During the research period, the bottom 10% group showed a response rate around 80%, the middle 80% group showed a response rate around 75%, and the top 10% group showed a response rate around 70%. Given that income is highly correlated with both response rates and expenditures, this variable was added to the nonresponse weighting process.

An analysis comparing the set of nonresponse variables from 2013 and prior to those for 2014 and forward showed the newer variables (which included IRS AGI) made some improvement in stratifying the sample households by response propensity, but they did not have a significant impact on raising the expenditures estimates, which is something CE was hoping for.⁶ This finding prompted additional research.

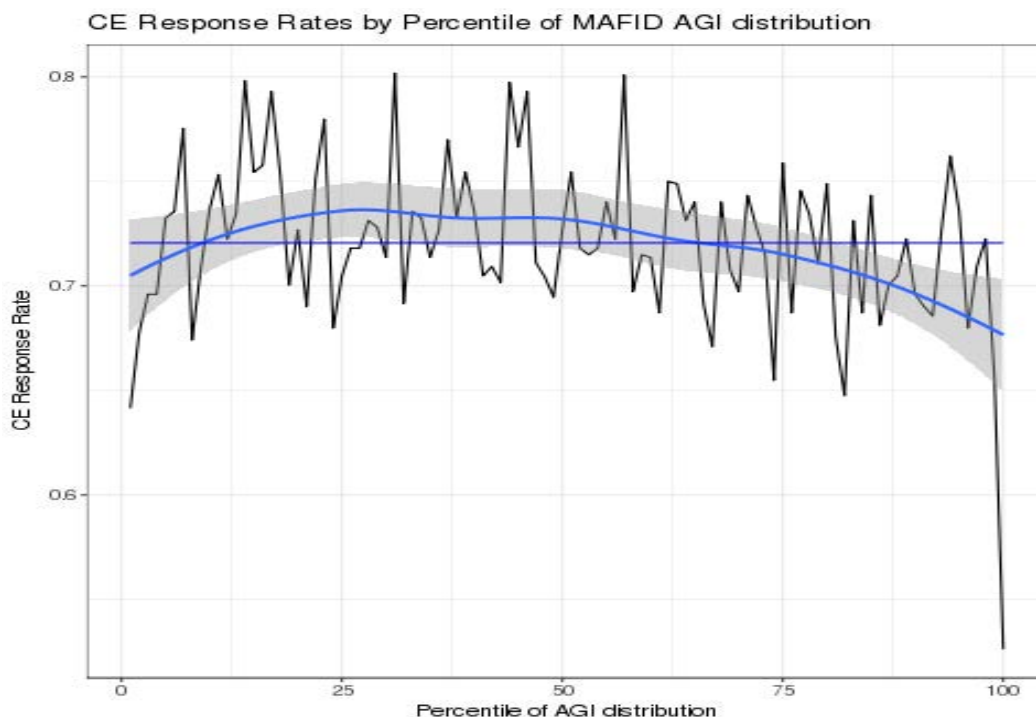
4. Exploring Data Linkage

As stated above, the new variables worked well in some aspects, but at the same time data linkage opportunities became available for household surveys to improve their weighting

⁶ Final Reports of the Survey Updating Monitoring (SUM) (2019) Team, Sharon Krieger, Brett McBride, and Taylor Wilson, pp. 15-29.

procedures by linking their own data to administrative data collected by other government agencies. One such opportunity became available for CE to link to IRS household data using AGIs from actual individual Form 1040 tax returns as opposed to the average AGI by zip code which CE had been using. The basic idea was that CE would continue using the same three income groups to stratify households in its nonresponse adjustment process, just replacing IRS's publicly available zip code level data with IRS's non-public household level data. Figure 2 below shows that CE Interview Survey response rates using IRS household level (Form 1040) income percentiles exhibit a similar distribution of response rates to the AGI zip code level income, especially at higher incomes.

Figure 2. CE Interview Survey Response Rates Using IRS Household Level (Form 1040) Income Percentiles



A major benefit of using household-level data is that it provides a more accurate estimate a household's income, which is valuable information under the assumption that a household's decision to respond or not respond to the survey is a function of its own individual income and not that of the community around it. These household-level data are available using the U.S. Census Bureau's data linkage infrastructure which is permitted under U.S. Title 13 and Title 26 for research purposes.

Two types of record linkage from the U.S. Census Bureau's infrastructure were used to match CE data to the IRS data.⁷ The first type of linkage was a housing-based linkage which used address and location information. Master Address File identification numbers (MAFIDs) were assigned to survey households and administrative tax units based on the available address information. The Master Address File is the Census Bureau's official list

⁷Quentin Brummet, Denise Flanagan-Doyle, Joshua Mitchell, John Voorheis, Laura Erhard, and Brett McBride, CARRA Working Series Paper 2018-01, pp. 4-8.

of all residential addresses in the U.S. The second type of linkage is a person-based linkage which used the Census Bureau's personal validation system. Protected Identification Keys (PIKs) were assigned to survey and administrative records and were based on personally identifiable information. For this research, only the MAFID address matching was used to link the CE data to the IRS data.

After the matching process was completed, issues like duplicates and other anomalies were addressed and the data were cleaned and deemed ready for use. The resulting data set showed nearly 70% of CE households were matched to at least one tax return with an adjusted gross income (AGI). Sometimes an address was matched to more than one tax return, and when that happened the AGIs from all the tax returns were summed together (up to a maximum of six) to represent the address's total AGI. Addresses with more than six tax returns were treated as non-matches since they were suspected of being apartment buildings or addresses with data linkage problems and not individual households. Of course, a match rate of 70% means 30% of the households still needed to have an income value imputed for them, and the imputation was done by assigning them the average AGI for their zip code using IRS's publicly available data as is done in CE's production weighting process.

5. Reclassification

After the matching and imputing was completed, the households were now ready to be reclassified into weighting groups based on the revised income variable. Table 1 below shows the number of households in the three income groups before and after reclassifying them by their household-level AGI values.

Table 1. Weighting Class Status Before and After Reclassification

		Income class after reclassification (using IRS's household-level income data)			Total
		Bottom 10%	Middle 80%	Top 10%	
Income class before reclassification (using IRS zip code level income data)	Bottom 10%	550	3,000	80	3,630
	Middle 80%	2,900	24,000	2,300	29,200
	Top 10%	250	2,200	1,200	3,650
	Total	3,700	29,200	3,580	36,480

This table shows before the reclassification process 10% of the households were in the bottom income group, 80% were in the middle-income group, and 10% were in the top income group. Then after the reclassification process 10% were still in the bottom income group, 80% were still in the middle-income group, and 10% were still in the top income group. Although the overall percentages stayed the same, many individual households changed their income groups. Looking at the diagonal cells it can be seen that 25,750 (= 550 + 24,000 + 1,200) of them stayed in the same income group after the reclassification process, which means 10,730 (= 36,480 – 25,750) of them changed to a different income group after the reclassification process.⁸

The reclassification process kept 70% of the households in the same income group and changed 30% of the households to a different income group. That is a fairly high

⁸ The numbers in this table are rounded due to confidentiality issues in accordance with U.S. Census Bureau Disclosure Review Board guidelines. Because of rounding, there are 36,480 households in this table as opposed to 36,638 stated earlier.

percentage of households that changed income groups. One way of seeing it is by considering the fact that by definition 10% of the households are in the low-income group, 80% are in the middle-income group, and 10% are in the high-income group, which means the maximum percentage of households that can change income groups is 40%. The fact that the percentage of households that actually changed income groups is close to the maximum percentage of households that can change income groups suggests that household incomes within zip codes are fairly heterogeneous, not homogeneous. High-income zip codes have a fair number of low-income households, and low-income zip codes have a fair number of high-income households. These results show that linking household level IRS income data was effective in assigning sampled addresses into more homogenous income groups for the nonresponse adjustment.

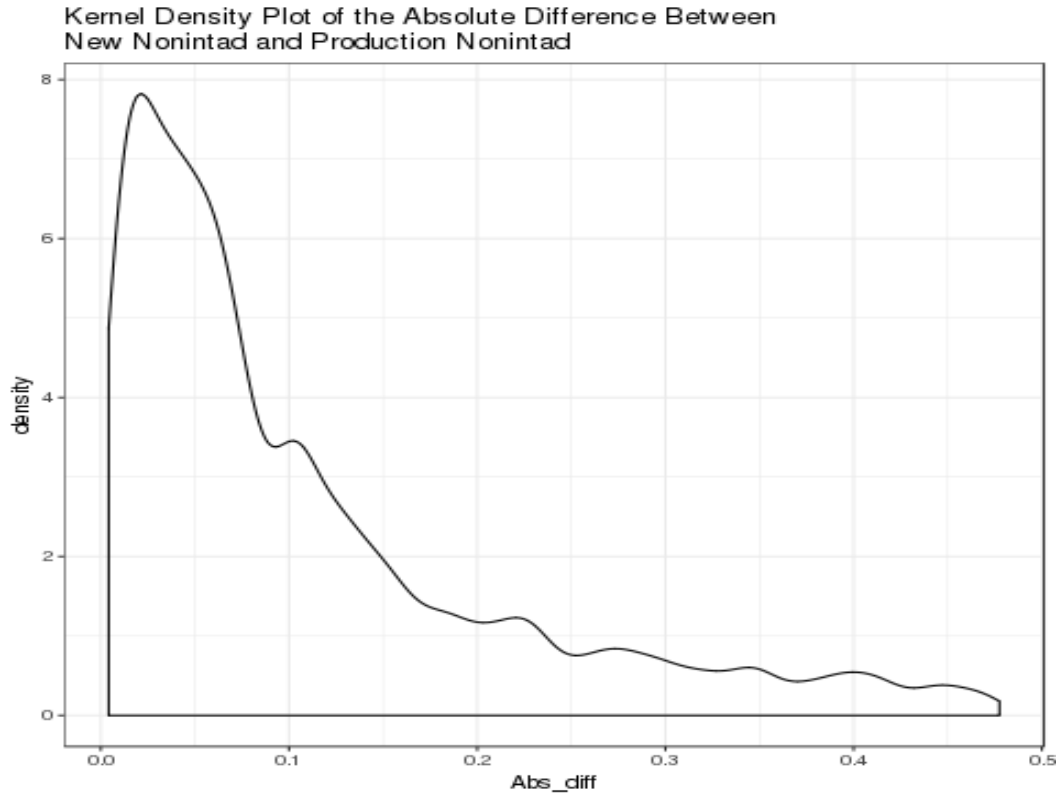
6. Reclassification Effect on Weights

After running the reclassified data through the nonresponse weighting process, the impact on the nonresponse adjustment factors and the final weights was noticeable. The average nonresponse adjustment factor in 2014 was 1.50, which corresponded to a response rate of 67% ($1.50=1/0.67$), and the average absolute value of the change in the nonresponse adjustment factors resulting from income reclassification was 0.13. The absolute value of the change was substantial and more than simply rounding. However, the actual nonresponse adjustment factor after reclassification was practically identical to the original value, around 1.50 as expected.

CE's "final weights" are the survey's base weights multiplied by a nonresponse adjustment factor and a calibration adjustment factor. Calibration adjustment factors make the survey's weights add up to certain "known" population totals, but they are not the topic of this paper so they will not be discussed here. The average final weight in 2014 was 19,600, which means each respondent household in the survey represented 19,600 households in the U.S. population, itself plus 19,599 other households that were not selected for the survey or were selected for the survey but did not participate in it. The average absolute value of the change in the final weights was 1,700, which again was substantial and more than simply rounding.

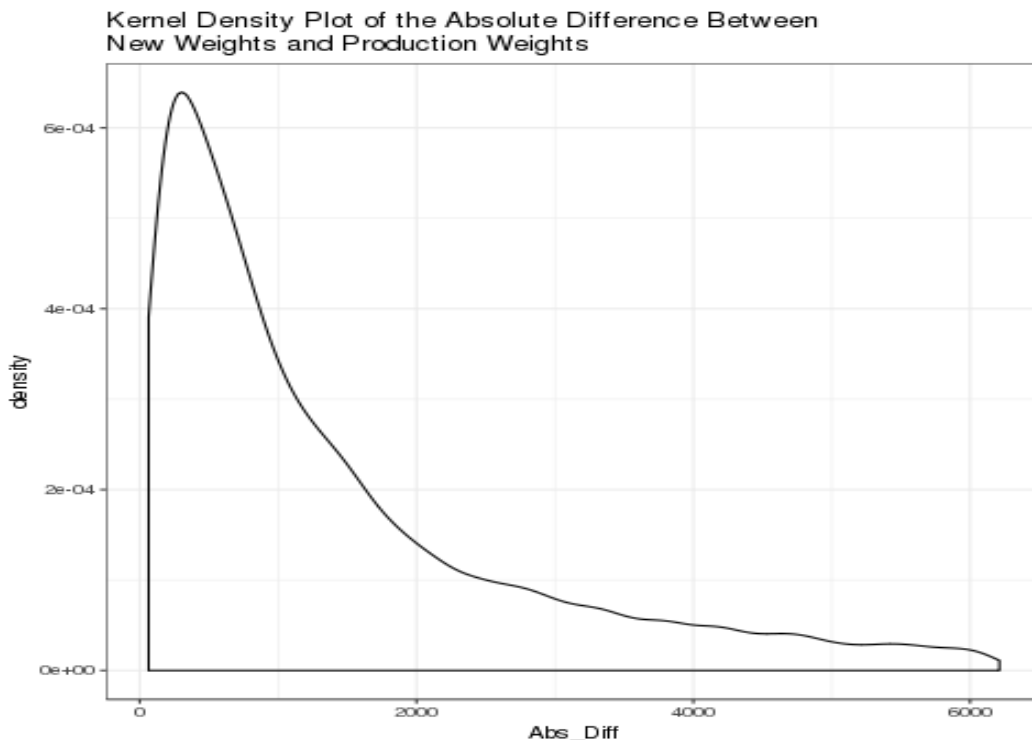
The distribution of the absolute changes in the nonresponse adjustment factors is shown in Figure 3 below as a kernel density plot. The absolute changes range from 0.00 to 0.50 with an average value of 0.13. The distribution is skewed to the right, so most of the absolute changes are smaller than the average value of 0.13.

Figure 3. Absolute Value of Difference between Production Nonresponse Factor and Reclassified Nonresponse Factor



The distribution of the absolute changes in the final weights is shown in Figure 4 below as a kernel density plot. It is similar to Figure 3 above. The absolute changes range from 0 to 6,000 with an average value of 1,700. Again, the distribution is skewed to the right, so most of the changes are smaller than the average value of 1,700.

Figure 4. Absolute Value of Difference between Production Final Weights and Reclassified Final Weights



7. Reclassification's Effect on Expenditures

One of the most important research objectives regarding the use of IRS household level data in the CE weighting process was its impact on expenditures. In academia, many believe that CE expenditures are under-reported⁹ so an overall increase in expenditures would be considered a positive finding and a major goal of this research. There were three quarters of data from collection year, 2014, used in this analysis, which included total expenditures and some of the major subgroup expenditures. In general, the results at the national level show that expenditures moved very slightly higher after reclassification and running the entire weighting process to calculate final weights. Specifically, for overall Interview Survey expenditures (ZOTALX4), there was almost no difference in expenditures between the means for 2014 quarter 2, around a 0.25% differential in 2014 quarter 3 with the reclassified weighting class means being higher and nearly a 0.5% differential in 2014 quarter 4 with the reclassified weighting class means being higher. The standard errors were practically identical between the production means and the reclassified for all three quarters of research.

⁹ John Sabelhaus, David Johnson, Stephen Ash, David Swanson, Thesia Garner, and Steve Henderson, *Is the Consumer Expenditure Survey Representative by Income?*(NBER Working Paper No. 19589, October 2013), pp.1-2.

Tables 2-4 displays the 2014 quarter 2 through quarter 4 comparison between CE's production mean and the means after reclassification at the national level. As mentioned above, there was almost no difference in 2014 quarter 2 but there was a rise to nearly 0.5% (0.46%) increase in ZTOTALX4 after running the reclassified data. All of the other major summary variables show a small increase, all less than 1% increase.

Table 2. 2014 Quarter 2 Comparison of Production Means to Reclassified Means at the **National** Level

Expenditure Variable	Production Mean	Reclassified Mean	Difference	Percentage Difference
All items	\$48,640	\$48,640	\$0	0.00%
Housing	16,584	16,570	-14	-0.08
Transportation	8,319	8,307	-12	-0.15
Food	7,631	7,624	-7	-0.09
Insurance	4,869	4,890	20	0.41
Health	3,942	3,955	13	0.31
Income before taxes	53,967	53,910	-58	-0.11

Table 3. 2014 Quarter 3 Comparison of Production Means to Reclassified Means at the **National** Level

Expenditure Variable	Production Mean	Reclassified Mean	Difference	Percentage Difference
All items	\$50,208	\$50,330	\$122	0.24%
Housing	16,739	16,760	21	0.13
Transportation	9,142	9,169	27	0.29
Food	7,648	7,669	21	0.27
Insurance	4,809	4,846	37	0.76
Health	4,212	4,211	-1	-0.03
Income before taxes	54,180	54,540	360	0.66

Table 4. 2014 Quarter 4 Comparison of Production Means to Reclassified Means at the **National** Level

Expenditure Variable	Production Mean	Reclassified Mean	Difference	Percentage Difference
All items	\$51,023	\$51,260	\$237	0.46%
Housing	17,155	17,210	55	0.32
Transportation	8,983	9,027	44	0.49
Food	7,660	7,681	21	0.27
Insurance	5,035	5,082	47	0.93
Health	4,198	4,230	32	0.76
Income before taxes	55,747	56,270	523	0.94

At the Census Region level, there were similar pattern for the means of total expenditures (ZTOTALX4) resulting from the reclassification as compared to the national level. As shown in Tables 5-7, all four regions saw a small increase or decrease but the differences were generally less than 1%.

Table 5. 2014 Quarter 2 Comparison of Production Means to Reclassified Means at the **Regional** Level for “All Items”

Census Region	Production Mean	Reclassified Mean	Difference	Percentage Difference
Northeast	\$53,513	\$53,690	\$177	0.33%
Midwest	46,699	46,910	211	0.45
South	45,615	45,550	-65	-0.14
West	51,604	51,390	-214	-0.42

Table 6. 2014 Quarter 3 Comparison of Production Means to Reclassified Means at the **Regional** Level for “All Items”

Census Region	Production Mean	Reclassified Mean	Difference	Percentage Difference
Northeast	\$56,396	\$56,080	\$-316	-0.56%
Midwest	46,636	46,520	-116	-0.25
South	47,374	47,620	246	0.52
West	53,477	54,040	563	1.05

Table 7. 2014 Quarter 4 Comparison of Production Means to Reclassified Means at the **Regional** Level for “All Items”

Census Region	Production Mean	Reclassified Mean	Difference	Percentage Difference
Northeast	\$58,480	\$58,680	\$200	0.34
Midwest	47,675	48,060	385	0.81
South	47,233	47,500	277	0.59
West	54,745	54,810	65	0.12

8. Future Considerations

In an effort to upgrade the nonresponse weighting process, this research focused on the potential improvement of CE’s current income variable. However, it is possible that there are additional administrative data variables available that may better capture a household’s propensity to respond and being correlated with expenditures. Perhaps a similar household-level wealth variable such total assets or assessed housing value could be a better alternative. It is even possible that environment is better at determining a household’s propensity to respond than personal wealth. For example, would living in a wealthy neighborhood at the block or tract level, regardless of your financial situation be more associated with response rates? All of these issues could be analyzed to determine if another variable would be a better candidate for the CE’s nonresponse weighting process.

9. Conclusion

The current CE weighting variables work fairly well but using actual IRS household-level income data in place of zip code level income did improve the current method as many households changed income groups through reclassification and were assigned to more homogenous income groups for the nonresponse adjustment process. This was made possible by successfully matching nearly 70% of CE households to actual IRS Form 1040 records. As a result of the successful matching of IRS records with CE data, the weighting process was run using the reclassified data to calculate new expenditure amounts.

However, the increase in these expenditures were not significantly higher. Any rise in expenditures is considered moving in the desired direction but the rise was not significant.