PWBM WORKING PAPER SERIES

MATCHING IRS STATISTICS OF INCOME TAX FILER RETURNS WITH PWBM SIMULATOR MICRO-DATA OUTPUT

Jagadeesh Gokhale

Director of Special Projects, PWBM
jgokhale@wharton.upenn.edu

Working Paper 2018-1
https://budgetmodel.wharton.upenn.edu/papers/2018/2/6/w2018-1

PENN WHARTON BUDGET MODEL
220 South 40th Street, Suite 250
Philadelphia, PA 19104
February 2018

The author is grateful to John Ricco and Nirvan Sengupta for excellent research and code-review assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the Penn Wharton Budget Model.

ABSTRACT

The PWBM Simulator implements micro-level projections of individuals and families in the United States calibrated to observed historical trends and interactions among many demographic and economic variables. The estimates and projections are rigorously validated using many sources of micro-data information in the United States. Using PWBM-Simulator output to implement tax policy analysis requires mapping its distributions of individuals and families into distributions of tax filing units with appropriate income elements calibrated to observed features of U.S. tax filers. This paper describes the procedures used for augmenting PWBM Simulator's micro projections with tax variables from U.S. tax returns reported in the Internal Revenue Service's public use Statistics of Income surveys.

Introduction

The Internal Revenue Service's Statistics of Income (SOI) section provides public use files containing information on individual income tax returns filed by taxpayers for a particular year. The files are released purely for research purposes. The latest public use file available at the time of this writing contains information for tax-year 2010 - hereafter called **SOI2010_PUF**. It is used to augment, calibrate, and appropriately reconfigure data generated by the Penn Wharton Budget Model Simulator to analyze the recently enacted tax reform in the United States. This paper describes the methods used in preparing and extracting information from **SOI2010_PUF** for that purpose.

Penn Wharton Budget Model is a micro-simulation calibrated to replicate in the computer, demographic and economic changes observed in the U.S. economy since the mid-1990s, and to use the observed momentum of demographic and economic forces to make projections of the population's composition, economic outcomes for population subgroups and for the aggregate U.S. economy. **SOI2010_PUF** micro-data information on U.S. tax filing units is used to distinguish PWBM Simulator families and individuals into tax-filing and non-filing units. In addition, **SOI2010_PUF** tax variables are used through a conditional assignment procedure to augment PWBM Simulator data for years 2010-2040. The resulting dataset on PWBM microsimulation tax-filing units during that period is the basis for PWBM tax-simulator and calculations of the likely budget and economic effects of the tax reform enacted at the end of 2017.

SOI2010_PUF contains 159,791 records, which is a sample of the 142.9 million Form 1040, Form 1040A, and Form 1040EZ Federal individual income tax returns filed during that year. Some of these returns are for years before 2010, but those are retained in the analysis as they represent returns generating revenues for the year 2010. The public use version of the tax-filer sample does not include Social Security Numbers (SSN), and other similar information to ensure

that no personally identifiable information is released publicly. To ensure tax-filer privacy, however, several additional masking procedures are also implemented on tax variables, making the file as released unsuitable for tax policy analysis. That's because masking procedures may modify or otherwise distort the distributional features of many important tax variables. Because the **SOI2010_PUF** is the only source of public information on income taxes and related variables, it is necessary at first to implement "unmasking" procedures to impute missing tax variables to prevent exclusion of important segments of tax-variable distributions from the analysis. These imputations are based on a set of assumptions designed to "guess" the missing distributional segments of masked tax variables based on the distribution segments that are observed in the sample. Making such assumptions is standard procedure in the use of **SOI2010_PUF** for tax analysis. The primary use made of **SOI2010_PUF** files is to provide general statistical tabulations relating to sources of aggregate incomes and taxes paid by individuals as well as to estimate the administrative and revenue impact of tax law changes.

The original **SOI2010_PUF** file that PWBM received from the IRS includes masked values of several variables and it does not contain demographic variables. Several masking procedures are employed (see Appendix 1) to blur or otherwise hide potentially identifiable information, including the values of several tax items reported by tax filer in their returns. Upon special request, the IRS provided PWBM staff with a supplement file, **SOI2010_SUP**, with demographic information on tax filer records via six additional variables. These variables include the age range of the primary filer (AGERANGE), ages of up to three dependents (AGEDP1, AGEDP2, and AGEDP3), the split of earnings between primary and secondary filer among joint filers (EARNSPLIT), and the gender of the primary filer (GENDER). **SOI2010_SUP** also contains the same ID variable – RECID – for matching observations in **SOI2010_PUF**. Unfortunately, only about two-thirds of tax filer observations in **SOI2010_PUF** are matched by observations in **SOI2010_SUP**.

This document describes the use made of the **SOI2010_PUF** and **SOI2010_SUP** files to generate federal individual income tax estimates and projects in PWBM Simulator. It describes

A. Procedures to impute missing/masked values in the merged PUF and SUP datasets: **SOI2010**

B. Procedures to assign non-wage income variables to individual tax filers in PWBM's PIT Simulator

C. Implementation of a SIM-SOI matching procedure to allocate SOI tax variables to PWBM Simulator filers.

Before describing these procedures, it is worthwhile to explore the potential for achieving a good match between SOI filer records and PWBM Simulator family records from which the subset of tax filers would be determined. Since the PWBM Simulator is closely calibrated to "agree" with CPS cross-section distributions along many demographic and economic outcomes, it's useful to compare SOI demographic and tax-variable distributions with those in the CPS.

The match potential depends on how closely the demographic composition of the filer populations approximate each other in the two data sources. Unfortunately, the **SOI2010_PUF** provided by the IRS does not contain demographic information for all observations. "Unmasking" procedures are used (see below) to impute missing values for several demographic variables such as age, gender, number and ages of dependents, and earnings split between joint filers. However, the **SUF2009_PUF** acquired by PWBM earlier has a full complement of demographic information.[1] Figure 1 shows the quality of the match between CPS and **SOI2009_PUF** distributions by demographic groupings -- multiple tax-filer groups distinguished by filer type (single or joint), gender (for single filers), number of children (0, 1, 2, 3+), and age (15-24, 25-34, 35-44, 45-54, 55-64, and 65+). Panel A of the figure shows the

[1] It was decided not to use the **SOI2009_PUF** for drawing tax variables since 2009 was a recession year.

distributions from the CPS and Panel B from the SOI. Figure 2 shows cross-correlations across the 10 income items in the CPS and SOI surveys.

Insert Figure 1.

One can also compare the distributions of total incomes from the two survey data sources. Income items that are common to the two surveys include wages, interest, dividends, rent, unemployment insurance, pensions, farm, business, social security, and alimony. Total income is first calculated as the sum of these ten income items for each individual in both surveys. This depends on how closely the distribution of total income in the CPS approximates that in the SOI, with total income calculating the ten available income items in both datasets (mentioned earlier). Unfortunately, the CPS is well known to seriously under-report several of the ten income items included in the analysis (wages, rent, dividends, interest, farm income, business income, pensions, alimony, Social Security, and unemployment benefits). However, cross-correlations across income items within the CPS appear to be very similar to those in the SOI.

Insert Figure 2.

The demographic and income cross-correlations in the two data sources suggests that a matching procedure based on demographics and income would work acceptably well for allocating tax variables from the SOI to individuals in the PWBM simulator.

A. Procedures to impute missing/masked values in the merged SOI2010_PUF data set

As a first step, a full data set is created by merging observations in **SOI2010_PUF** and **SOI2010_SUP.** The merging preserves all observations in **SOI2010_PUF.** Values of the six variables in **SOI2010_SUP**

that are not matched to those in **SOI2010_PUF** are initially set to "missing." These variable values are imputed using the procedures described below.

1. First, the analysis sample is selected by eliminating non-2010 returns and the "reference return" which summarizes (a very few) extreme-AGI valued observations that are masked in the **SOI2010_PUF** (see Appendix 1 for greater detail).
2. The number of dependents is calculated as NUM_DEP = XTOT – (XFPT + XFST), where XTOT is the total number of exemptions claimed. XTOT is top-coded at 5. XFPT is the primary filer exemption {0, 1} and XFST is the spousal exemption {0, 1}. It's never the case that XTOT=5 and NUM_DEP = 4. That means NUM_DEP is top-coded at three and that the number of dependents up to three is known for all SOI-2010 observations.
3. The FAMTYP variable is created to classify observations into 12 categories by (single-filer male with {0, 1, 2, 3+} dependents; Single female filer with {0, 1, 2, 3+} dependents; and joint filers with {0, 1, 2, 3+} dependents.)
4. Values are assigned for dependents' ages: AGEDP**X** variables (**X**=1, 2, or 3+) – the ages of up to three dependents for 2010 filers. It ensures that the same number of AGEDP**X** variables are filled in with non-zero values of the age variable as there are number of dependents (NUM_DEP) in the filer unit. Figures 3 provides examples of the cumulative distributions calculated from the 2010 SOI file with unmasked dependents' ages to accomplish the imputation for masked values. For single filers (Figure 3), the x-axis shows the ages of dependents and the different curves represent the age-range of single filers with just one dependent. Panel A of Figure 3 shows the example of single filers: It shows that when such filers are older, the probability that the dependent is older is higher. Panel B of Figure 3 indicates systematic variation in dependents' ages conditional on the number of dependents: For single filers aged 35-46,

with 1, 2, and 3+ dependents respectively, (dotted lines in Figure 2), Panel B shows that given filer age-range, the larger the number of dependents, the higher the likelihood of having younger dependents. Dependents' age CDFs for joint filers (unbroken lines in Panel B) show that the probability of having younger dependents is even larger than for single filers and it, too, increases with the number of dependents.

Insert Figure 3.

This leaves the variables AGERANGE, GENDER, and EARNSPLIT unchanged – that is, some observations still have these values unassigned. The masking of GENDER means the FAMTYP variable cannot be computed for all observations.

5. The identical procedure as in (a) above is used to assign AGEDP**X** values for masked observations in the **SOI2009_PUF** file.
6. Most straightforward would be to impute the masked values of the AGERANGE variable (that categorizes filing units by the primary filer's age) by constructing conditional distribution functions of AGERANGE from the observations for which AGERANGE is not masked in the SOI2010 PUF itself. It would be desirable to do this separately for each **family type** and by different levels of AGI (provided by the variable AGIR1). Family types are distinguished by whether they are single or married (married couples are assumed to be joint filers), by gender of single filers, and the number of dependents (0 up to 3) in the family. That makes 12 FAMTYPEs in all: Male single filers with 0-3 dependents, Female single filers with 0-3 dependents, and joint filers with 0-3 dependents. Combined with the six categories of AGERANGE (of family head, or primary filer), it makes 72 filer types (demographic groups) in all. Unfortunately, two of the three variables for constructing the family types (filer type and GENDER) are also masked in the

SOI2010_PUF. However, these two variables are not masked in the **SOI2009_PUF**. Hence, imputations for AGERANGE in the **SOI2010_PUF** are implemented using CDFs of AGERANGE by family type and AGIR1 derived from **SOI2009_PUF**. There are two steps involved:

a. First, masked AGERANGE values are imputed in **SOI2009_PUF** by first constructing CDFs of AGERANGE by family-type and AGI values (included in SOI surveys as variable AGIR1). Figure 4 shows AGERANGE CDFs for selected AGI ranges

Insert Figure 4

7. Non-financial variables from the resulting **SOI2010_PUF** and **SOI2009PUF** datasets are output to flat-files.
8. For each filer observation in **SOI2010_PUF** with given gender and where demographic variables are masked, differences are calculated across 35 common variables that describe filer characteristics in **SOI2009_PUF** and **SOI2010_PUF**. The differences are summed and squared. Finally, and the imputation of demographic variables (AGERANGE, and EARNSPLIT) is implemented from the observation in **SOI2009_PUF** with the smallest sum of differences with the **SOI2010_PUF** observation in question. **SOI2009_PUF**.
9. Once the GENDER, AGERANGE, and EARNSPLIT attributes have been imputed in this manner for all observations with missing data in the **SOI2010_PUF** file, the family-type variable can be determined and assigned. All of the non-financial variables are merged with all financial variables and output to a file **SOI2010^{+}_PUF**. This dataset is used to match and merge (bootstrap) SOI tax variables with the PWBM micro-sim filer observations augmented by income items from the CPS2011 survey's filer observations (see below).

B. Procedures to assign income values to tax filers in PWBM's Simulator

The PWBM Simulator assigns wage income to individuals aged 15 and older based on their demographic and economic attributes. To construct a fuller income picture for calculating income taxes, the first step is to assign non-wage income items to individuals in the SIM. Values for income variables available in CPS surveys are assigned to PWBM Simulator individuals after matching each PWBM Simulator family with a family in the corresponding year's CPS survey. Several steps are involved:

1. Prior to the matching, CPS person-level data are collected into family-level records (by family) wherein married joint filers wage and other incomes are added together. In each year between 2010 and 2016, CPS filer records are grouped by the same family-type categories as described earlier. The matching of PWBM Simulator and CPS filer records done within each of the 72 family groupings. Such matching is done for PWBM simulator data for years 2010 through 2016 (CPS Survey years 2010-2016).
2. The SIM-CPS matching includes an adjustment for top-coded income variables. Again, the adjustment is done within corresponding demographic groups, but this time the source of the imputations is the **SOI2010^{+}_PUF**. Within each demographic group, each PWBM Simulator observation's CPS income item that is at the top-coded value is replaced by an average of three random draws of the corresponding income item from the **SOI2010^{+}_PUF** out of those income observations that exceed the CPS top-code. This procedure is repeated for each of the 10 income items including income from wages, dividends, interest, rent, unemployment compensation, farm income, business income, pensions, social security, and alimony. Income items allocated for years after 2016 are inflated using a historical wage-index growth rate.

3. The next step in the matching process is to complete the number and ages of dependents. The PWBM Simulator and CPS provide information only on the number of children aged 17 and younger within families. However, dependents living outside the home – such as college students older than age 17 may be claimed as dependents on their parents' returns. This may be true regardless of whether older children earn income and themselves file tax a return. The number of such older dependents is imputed using conditional distribution functions by FAMTYPE and AGERANGE constructed from **SOI2010^{+}_PUF** (after dependent's ages been imputed therein for missing demographic information from **SOI2010_SUP**.) The conditioning variables are filer-types and the number of younger dependents in the filer unit.[2] These dependent assignments are implemented to PWBM simulator annual filer data for years 2010 through 2040.
4. Filer/non-filer status is determined The rate at which low-income families file tax returns is calibrated to information from studies and notes available on the IRS website. For example, only 80 percent of low wage earning families who are eligible for EITC are deemed to file tax returns.[3] The tax-filing rules and dependent claiming rules delineated in the IRS Form 1040 instructions are applied.
5. The completion of the above step generates two data sets: The **SOI2010^{+}_PUF**, and the PWBM simulator's annual complement of tax filing units for years 2010 through 2040, each distinguishing 72 different filer types.

[2] The maximum number of dependents reported in the **SOI2010_PUF** is 3, where "3" is interpretable as "three or more." The conditional distribution functions of the number of older dependents for filers with zero young dependents (aged 0-17) is defined over the range of 0-3 older dependents. For those with one younger child in the family, the CDFs for older dependents are defined over 0-2, and so on.

[3] The 80 percent filing rate employed in the assignment of families into filers and non-filers is based on a 2017 IRS memo available at: https://www.irs.gov/newsroom/the-earned-income-tax-credit-often-missed.

6. The final step is to assign to each PWBM simulator observation (in each year between 2010 and 2040), an observation from the **SOI2010^{+}_PUF.** Once the assignment is done, the tax variables are inflated by a historical average rate of wage growth. The assignment of observations is done on a weighted basis. Each **SOI2010^{+}_PUF** observation has an associated weight (provided by the source data set **SOI2010_PUF**) whereas PWBM simulator observations represent a miniature sample of the entire US population in that year and all filing units have the same weight equal to the scaling factor (described below). In order to span the full set of **SOI2010^{+}_PUF** observations (within each of the 72 filer types), the assignments are made using a rank-order matching technique following the method adopted by the Congressional Budget Office.[4] This technique requires creating additional observations in PWBM simulator data of filers (in each year in the range 2010-40) so that the sum of weights in the resulting dataset equals that of the source data set (**SOI2010^{+}_PUF**). The resulting data set (spanning years 2010-40) is called the PWBM^{+} simulator dataset.
7. After assignments of tax variables has been completed, the PWBM^{+} simulator data sets for 2010-40 can be used as the basis for individual income tax calculations. Note that the PWBM simulator data set already distinguishes filing and non-filing units and tax variables are assigned to all filing units. It is a complete data set as regards representing the economy-wide population of filing units. Hence, aggregate tax variables calculated based on filing units in the PWBM^{+} simulator dataset. Since it is a miniature representation of the economy-wide population, however, calculating tax aggregates involves the application of a

[4] See https://www.cbo.gov/publication/52914.

scaling factor. The scaling factor is calculated by taking the ratio of the total U.S. population in 2010 to the weighted total PWBM$^+$ population. Table 1 shows the quality of the match between the U.S. economy-wide totals of various income and revenue aggregates, and the corresponding aggregates calculated from PWBM$^+$ simulator dataset.

Item	US Economy Total	PWBM$^+$ total	Percentage difference
Wages	**5,847,248,809,083**	**5,754,274,595,406**	**0.984099**
Total Income	**8,553,883,279,745**	**8,327,682,843,334**	**0.973556**
AGI	**8,137,825,583,212**	**7,890,408,617,705**	**0.969597**
Deductions	**1,994,146,071,586**	**1,954,503,132,842**	**0.980120**
Taxable Income	**5,531,159,085,139**	**5,319,150,759,147**	**0.961670**
Tax Before Credits	**1,067,403,149,690**	**1,009,567,426,953**	**0.945816**
Credits	**197,375,661,879**	**193,474,774,634**	**0.980236**
Tax After Credits	**954,676,288,337**	**900,304,782,532**	**0.943047**
Schedule C Income	**267,453,895,274**	**261,537,291,621**	**0.977878**
Schedule D Income	**376,104,738,466**	**321,967,970,387**	**0.856059**
Schedule E Income	**453,789,701,558**	**428,808,206,414**	**0.944949**
Schedule F Income	**(11,391,553,411)**	**(11,230,965,750)**	**0.985903**

Table 1: Totals and Percentage Differences between SOI2010+_PUF and PWBM+ Simulator Income and Tax Items.

It should be noted that the advantage of the method described here for assigning tax variables from **SOI2010$^+$_PUF**, separately by filer type in each year of the PWBM simulator dataset is to preserve the effect of demographic changes as generated by the baseline demographic simulation.[5] Figure 5 shows projections of income and tax aggregates calculated from PWBM$^+$ simulator data, inclusive of projected demographic changes in the relative sizes of the 72 demographic groupings by filer type.

Figure 6 provides a visual comparison of wage distributions (selected percentile values of wages between the 10th and 90th percentile) calculated from the Current Population Survey alone

[5] See the PWBM website for additional detail on PWBM's demographic projection methods and outcomes.

(Panel A) and that calculated from PWBM-Simulator output based on matching observations from the Statistics of Income (Panel B) data. Each individual hump-shaped segment reflects the wage distribution by age categories within a particular family type. There are 12 family types: Single Females with, alternatively, zero, one, two, or three dependents; single males with, alternatively, zero, one, two, or three dependents; and joint filers with, alternatively zero, one, two, or three dependents. Figures 7, 8 and 9 show similar information about the wage distributions for progressively higher segments (percentile ranges) of those distributions.

The differences in wage distributions from the two data sources in Figures 6 through 9 indicate that the rank-order matching procedure "improves" the PWBM Simulator's wage distribution compared to simply using wage information from Current Population Survey. As is well known, wage and income values are under-reported in the Current Population Surveys with the under-reporting concentrated at high earnings levels. The improvement in the distributions purchased by adopting the matching procedure with SOI data is simply in terms of increasing wage values, especially at the upper reaches of the distribution.

It is well known that wage and salary responses in the CPS suffer from underreporting. The procedure adopted here is to first separate PWBM Simulator families into tax filers and non-filers based on filing requirements in the income tax code. That division is based on wage and self-employment income in the CPS. Given underreporting of such incomes, it is likely that the non-filer group is too large. Replacing CPS incomes with income imputed from SOI for filers would be expected to correct the under-reporting in that group. However, to prevent the base for payroll taxes from being under-stated, it is necessary to adjust non-filers' income (wage and self-employment income subject to payroll taxes) as well. The adjustment is based on calculating the residual shortfall and re-scaling non-filers' incomes by a constant percentage in each year of the

historical simulation. The average of such annual percentage adjustments is applied to years beyond 2010.

Conclusion

The PWBM Simulator provides micro-level projections of individuals and families in the United States calibrated to account for most major trends and interactions among demographic and economic variables. Using PWBM-Simulator's demographic projection base to implement tax policy analysis requires mapping its distributions of individuals and families into distributions of tax filing units with appropriate income elements calibrated to observed features of U.S. tax filers. This paper provides details of the procedures used for augmenting PWBM Simulator's micro projections with tax variables from the IRS's public use tax-return available in the Statistics of Income survey.

Appendix 1

SOI Procedures to Eliminate the Potential for Identifying Individuals Tax Filers

The masking procedures employed seek to preserve the character of the microdata file while also protecting the identity of individuals. The changes made to individual tax records include:

1. Fiscal year returns have been converted to reflect the most recent year-end Tax Year and returns older than five years (Tax Year 2006 or less) have been removed from the file.

2. Returns containing one or more amount fields with deemed extremely large values are excluded from the microdata sample and are aggregated into two records special ID codes for returns reporting extreme negative and extreme positive AGI values. Values are considered extremely large if they are, generally, within the highest 30 amounts reported for any income amount value or within the lowest 30 amounts reported for any negative income. In all, 1,155 tax returns are aggregated, representing 1,379 returns in the population.

3. Returns that were sampled as a high-income-no-tax return, at a rate of 100 percent, have been placed back in their regular strata based on total income and subsampled at the corresponding strata rate.

4. To ensure that it will be impossible to know whether a given taxpayer is represented in the sample, all remaining returns sampled at rates greater than 10 percent[3] have been subsampled at 10 percent.

5. Those records sampled at a rate greater than 0.07 percent have been altered so that:

    a. Alimony paid, alimony received, and the State sales tax deduction are removed.
    b. Marital status is modified.
    c. Personal exemptions are modified per phase-out limitations and moved to other items.
    d. Multivariate blurring is applied to these returns with nonzero values in at least two of the following fields: wages and salaries, state and local income taxes, and real estate taxes. Prior to blurring, these returns are grouped into one of 10 categories based on their filing status and the number of dependents, and then further grouped by the pattern of nonzero values on these three fields plus presence of Schedule C, which is used only for grouping. A multivariate distance statistic is then calculated from the nonzero values of the three variables within each group. Based on this statistic, the two most distant records are identified, and the two additional records closest to each of these two records are located. For each group of three records the average value of each variable is placed in the specific fields. This process is repeated until all records have been averaged or "blurred."

6. All returns sampled at a 0.07 percent rate have been blurred on a univariate basis for the following fields: alimony paid, alimony received, wages and salaries, medical and dental expenses, real estate taxes, and state and local income taxes. Alimony paid and alimony received are blurred nationally. Prior to blurring wages and salaries, the records are

grouped into one of 25 categories based on filing status, number of dependents, and sample code. Prior to blurring medical and dental expenses, the records are grouped into one of 8 categories based on age range of the primary and filing status. Prior to blurring real estate taxes, the records are grouped into one of 21 categories based on filing status, number of dependents, and sample code. Prior to blurring state and local income taxes, the records are grouped into one of 17 categories based on filing status, grouped sample code[4].

7. All returns filed with marital status "Surviving Spouse" have been converted to Married Filing Jointly.

8. For all records on the file, the total number of dependents is capped based on filing status. For joint and head of household returns the total number of dependents shown is capped at 3, for single returns the total number of dependents shown is capped at 2, and for married filing separately returns the total number of dependents shown is capped at 1.

9. For all records on the file, all amount fields have been rounded. Amounts, in absolute values, above $100,000 are rounded to the four most significant digits (e.g., $228,867 = $228,900 and $1,158,235 = $1,158,000). Amounts between $10,000 and $100,000 are rounded to the nearest $100. Amounts between $5 and $10,000 are rounded to the nearest $10. Nonzero amounts less than $5 are set to $2, with sign retained.

10. Finally, all records in the file are rebalanced to ensure accounting accuracy after the above disclosure procedures are applied. Since individual records in this file may or may not contain data from just one tax return – and never contain the full item content of any one tax return

These information-masking procedures imply that the SOI2010_PUF and SOI2010_SUP files do not represent individual income tax returns. However, they represent acceptably well the distributions of tax variables across the population of tax filing units (families and households) across the United States.

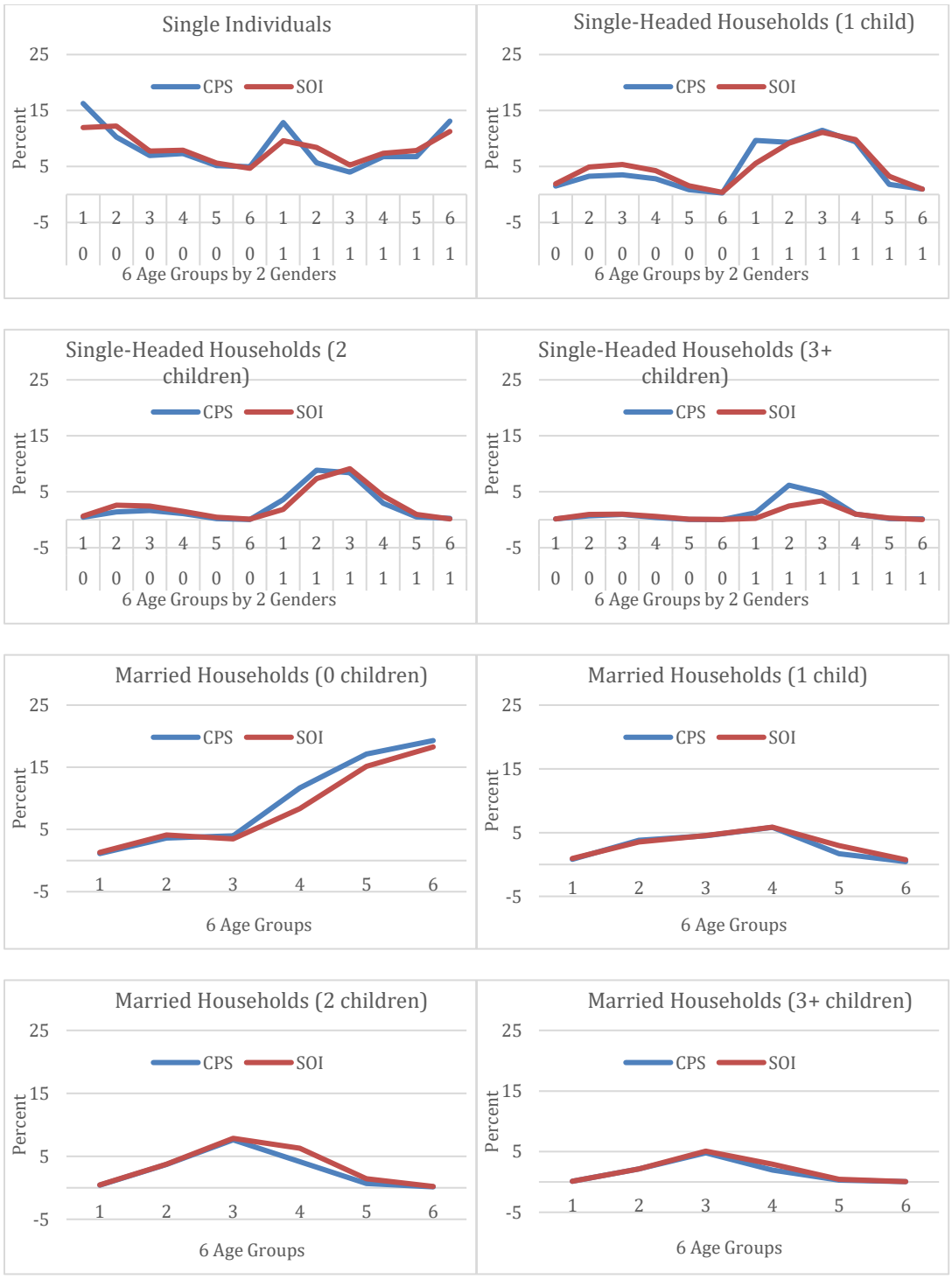
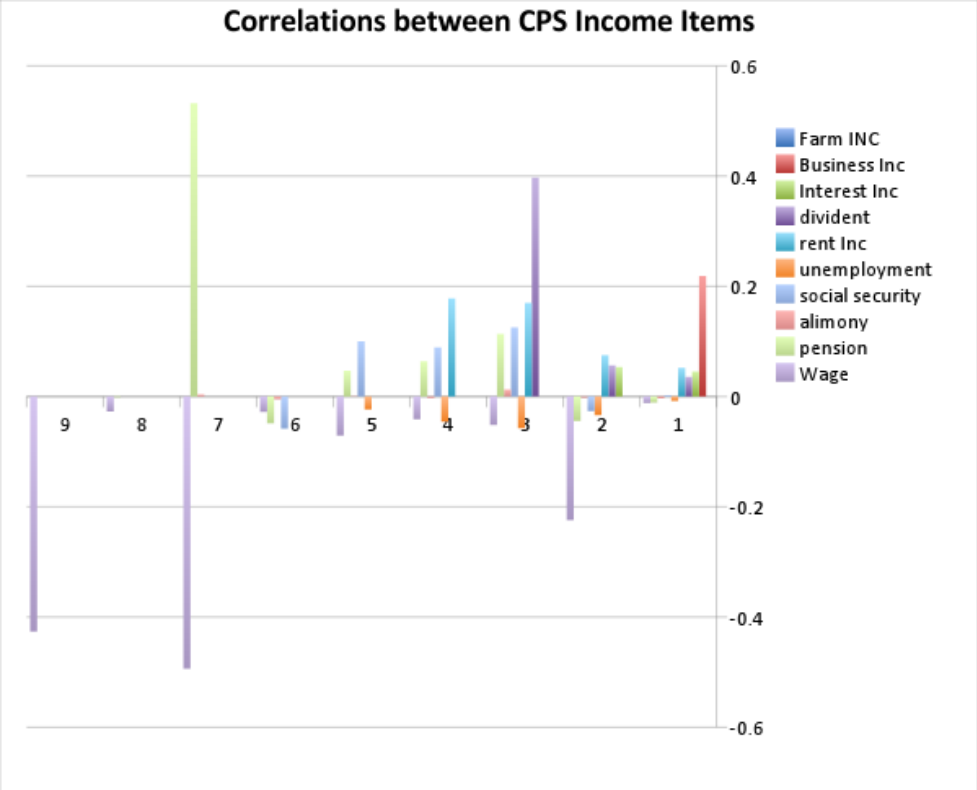


Figure 1: CPS and SOI Demographic Match of tax filing units.



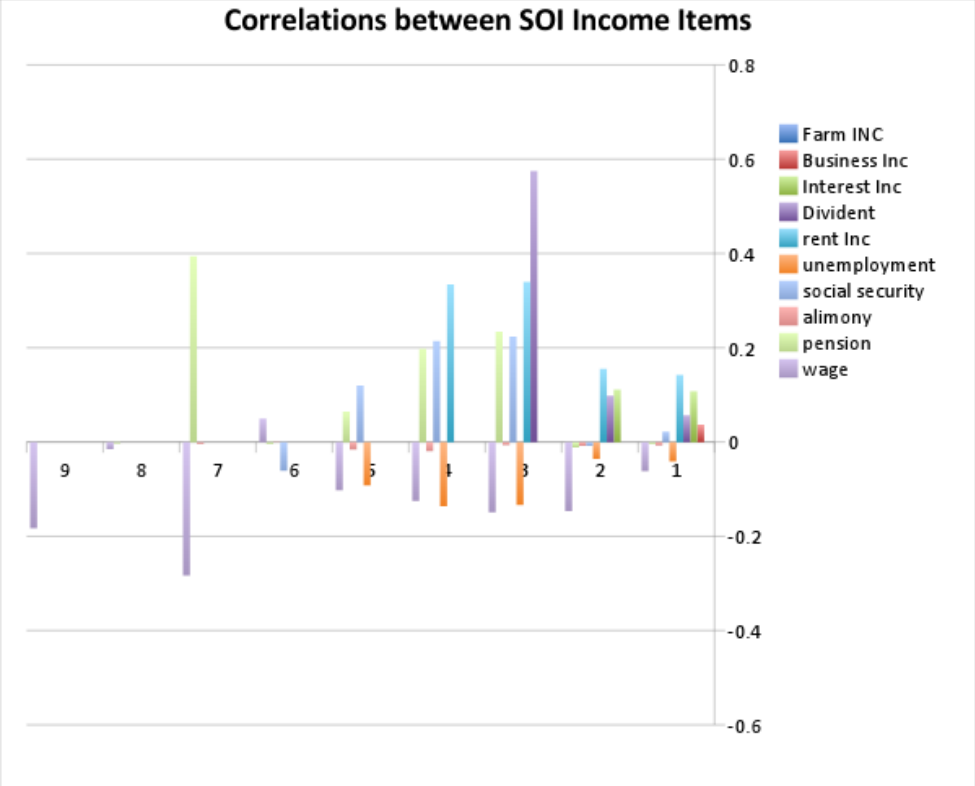
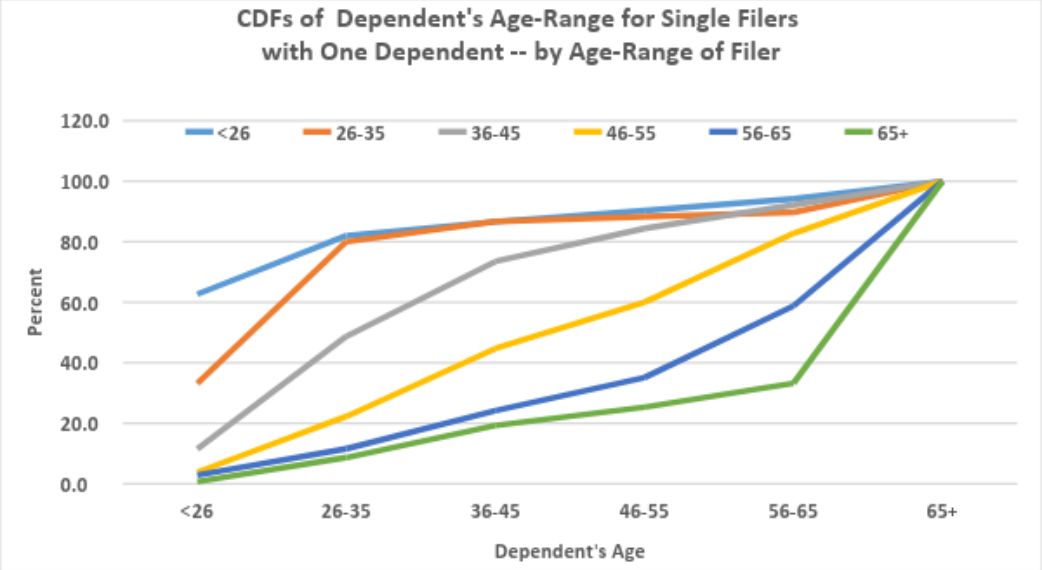


Figure 2: Correlations between Income Items in CPS and SOI.

Panel A



Panel B

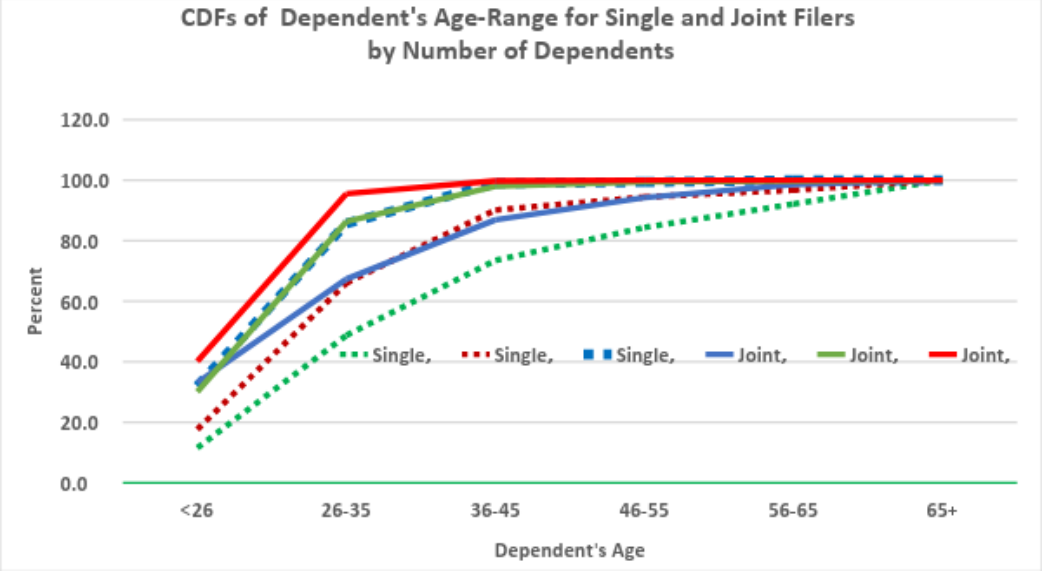


Figure 3: Conditional Distributions of Dependents' Age Range for Single and Joint Filers.

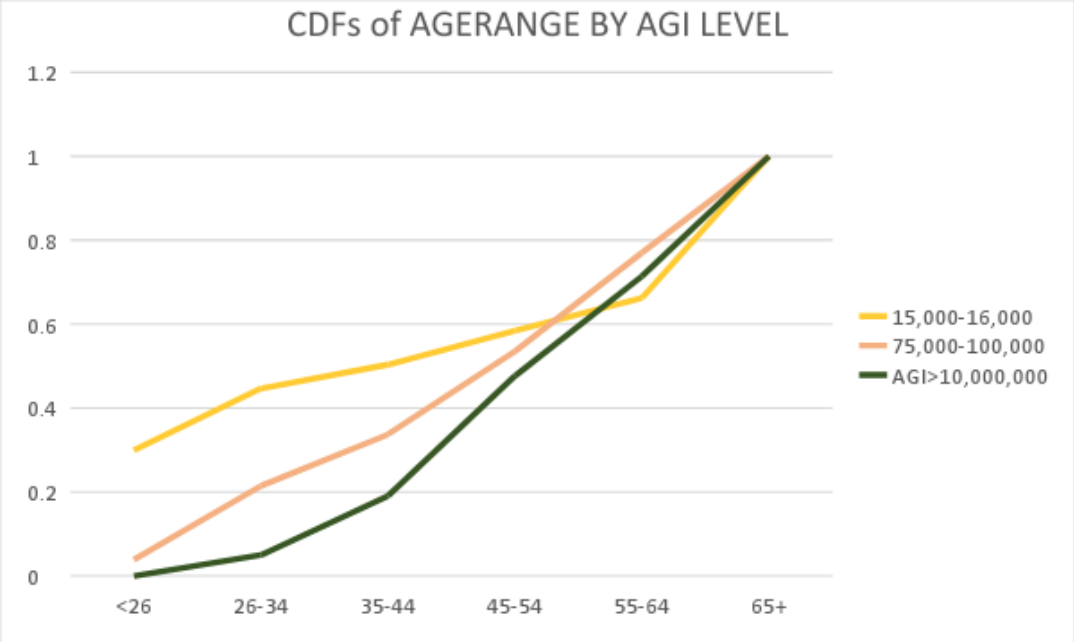


Figure 4: Estimated Conditional Distributions Functions of AGERANGE by Filer Age at selected AGI levels.

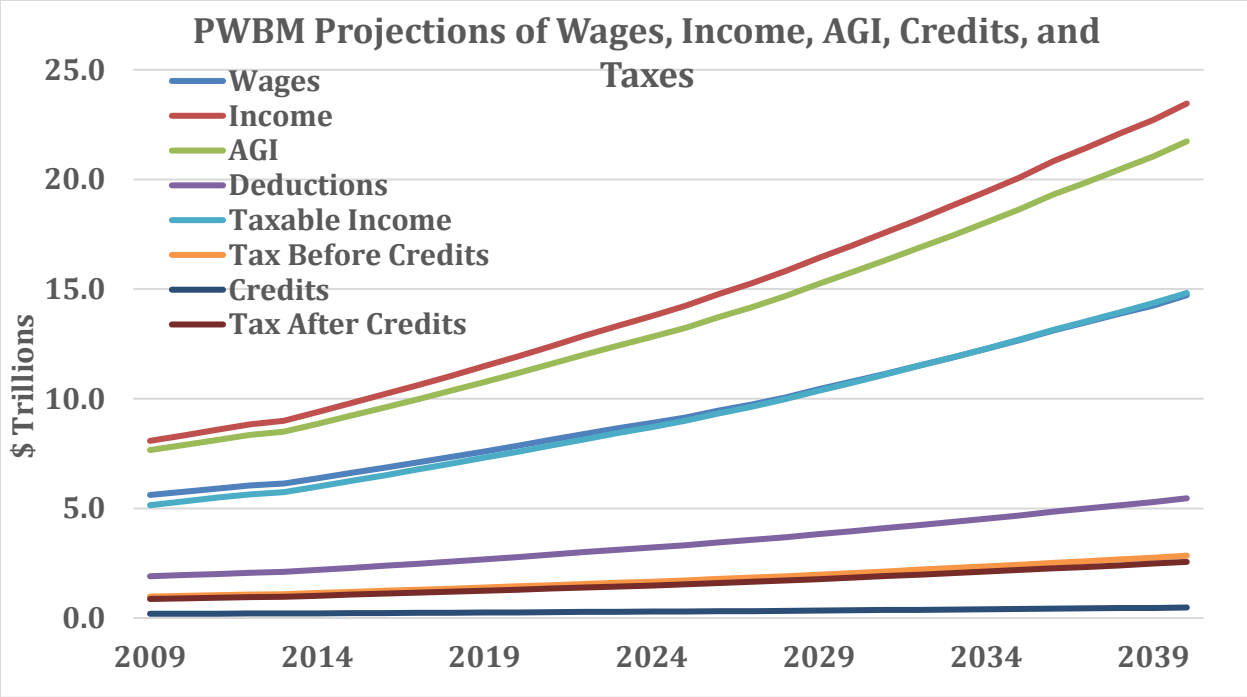
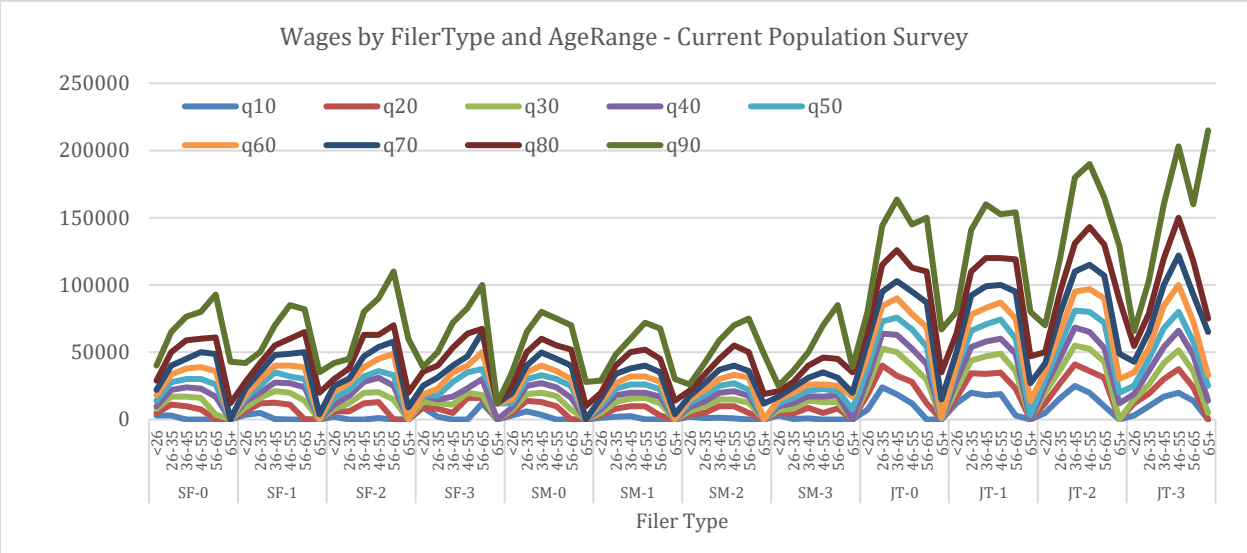


Figure 5: PWBM projections of Wages, Total Income, AGI, Credits, and Individual Income Tax Items.

Panel A



Panel B

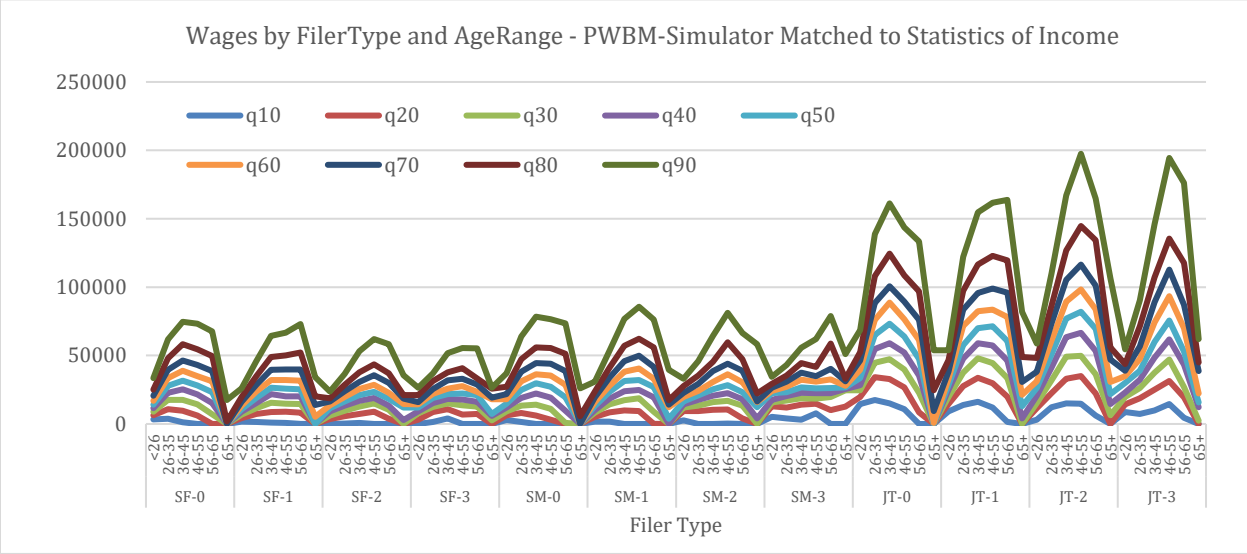
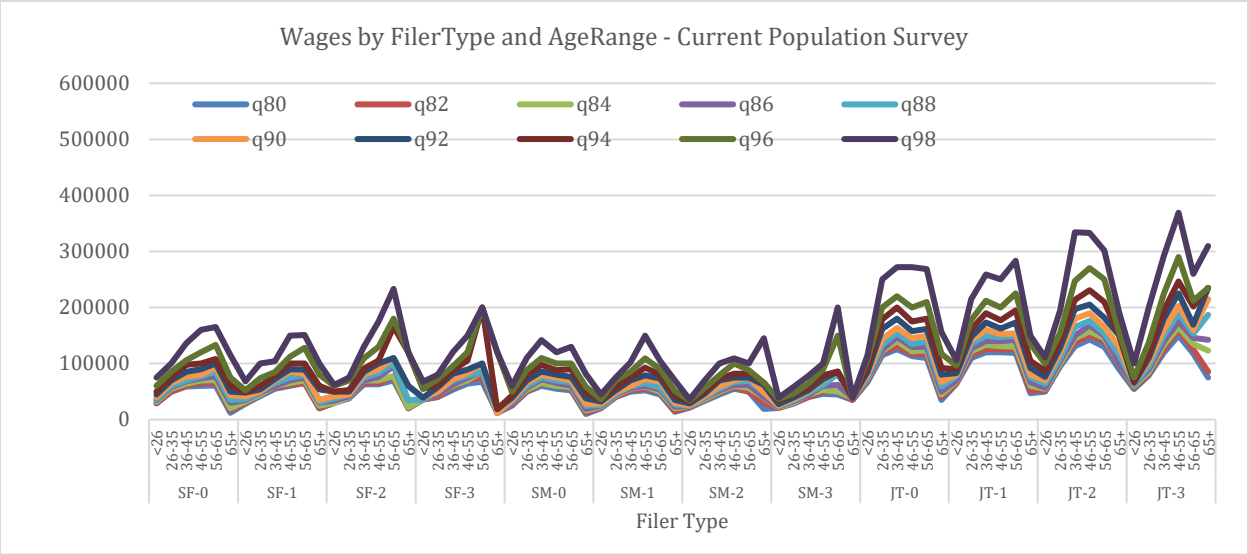


*Figure 6: Distributions of wage earnings within groupings by age, gender, filer-types, and number of dependents – **10th-90th percentile values** – comparing PWBM Simulator output with Statistics of Income 2010 data with Current Population Survey 2011 data.*

Panel A



Panel B

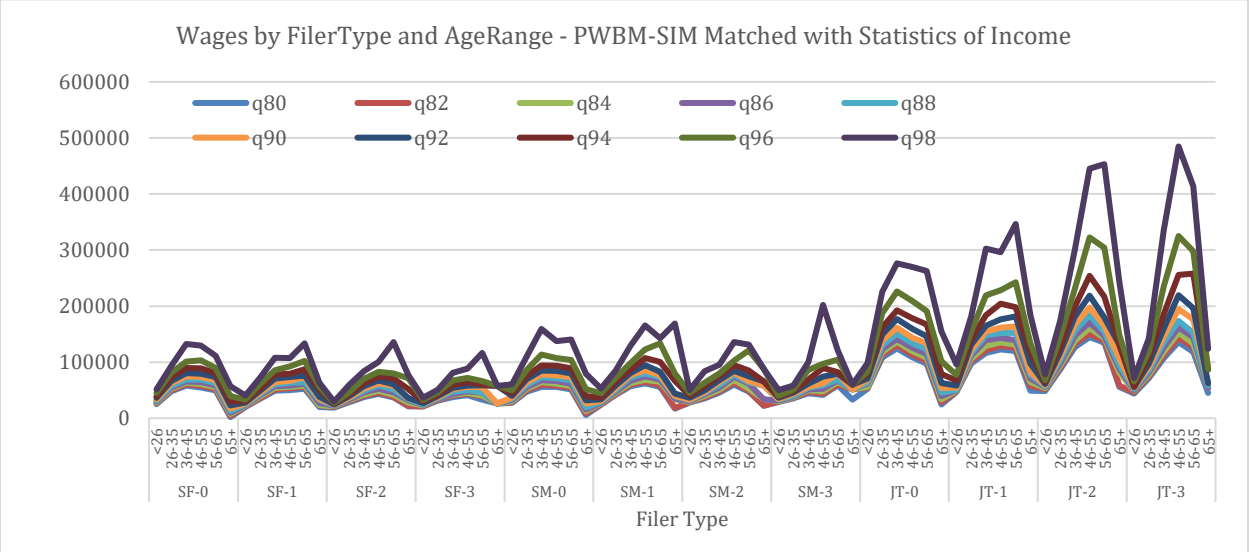
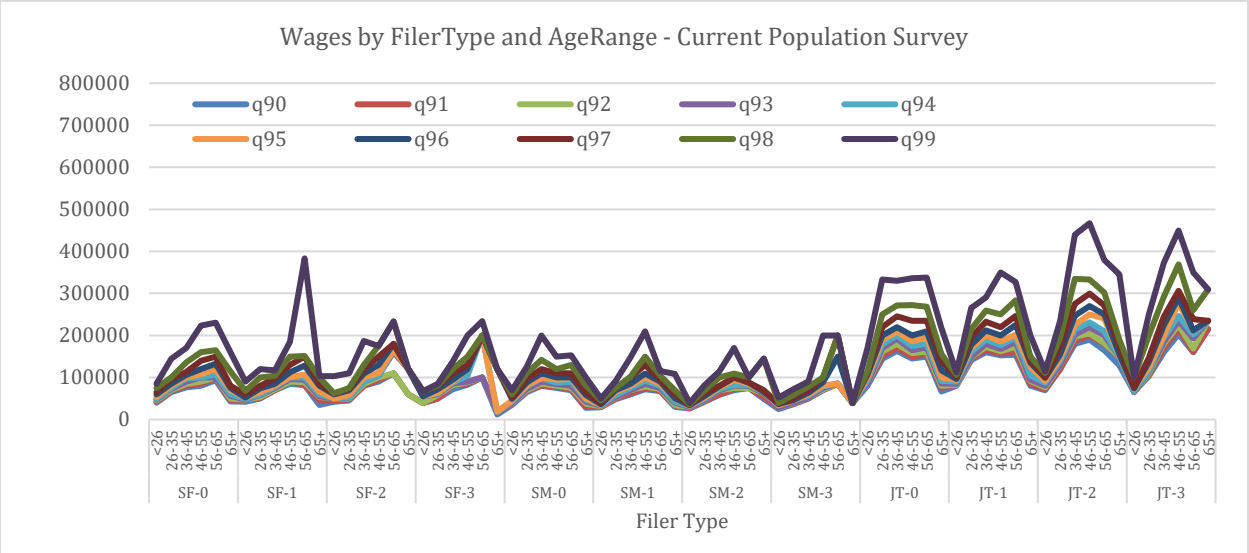


*Figure 7: Distributions of wage earnings within groupings by age, gender, filer-types, and number of dependents – **80th-98th percentile values** – comparing PWBM Simulator output with Statistics of Income 2010 data with Current Population Survey 2011 data.*

Panel A



Panel B

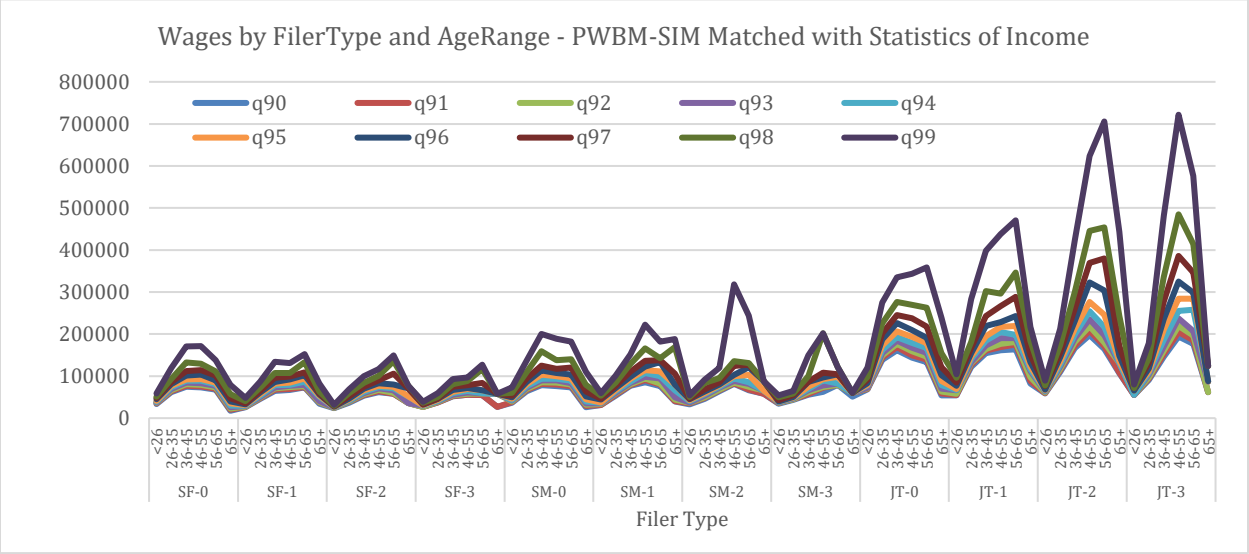
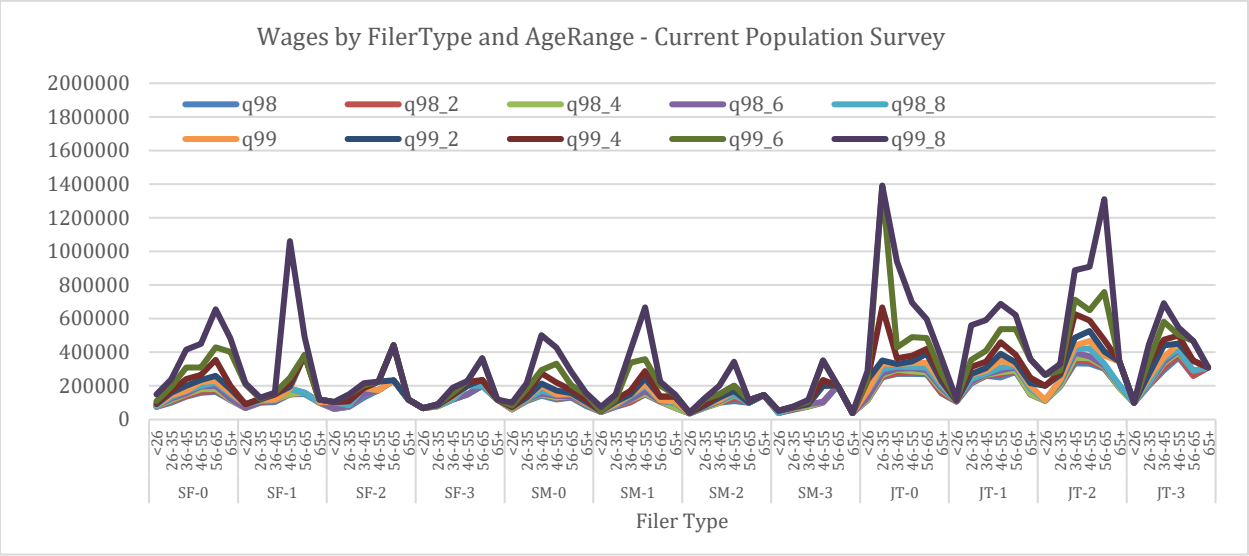


*Figure 8: Distributions of wage earnings within groupings by age, gender, filer-types, and number of dependents – **90th-99th percentile values** – comparing PWBM Simulator output with Statistics of Income 2010 data with Current Population Survey 2011 data.*

Panel A



Panel B

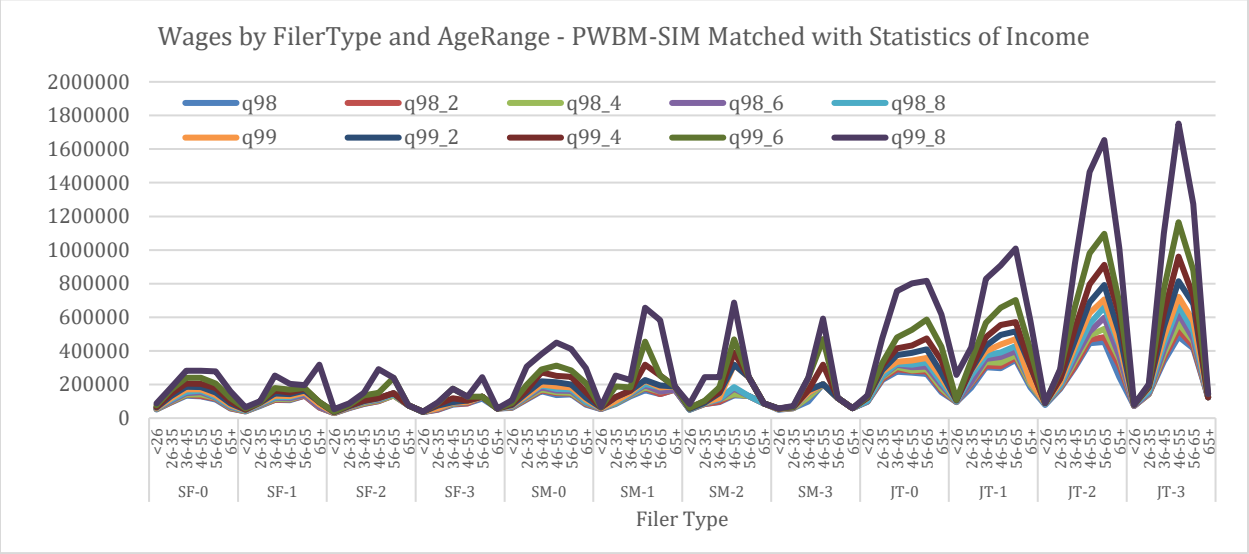


*Figure 9: Distributions of wage earnings within groupings by age, gender, filer-types, and number of dependents – **98th-99.8th percentile values** – comparing PWBM_Simulator output with Statistics of Income 2010 data with Current Population Survey 2011 data.*