Mobility, Opportunity, and Volatility Statistics (MOVS): Infrastructure Files and Public Use Data

by

Maggie R. Jones
U.S. Census Bureau

Adam Bee
U.S. Census Bureau

Amanda Eng
U.S. Census Bureau

Kendall Houghton
U.S. Census Bureau

Nikolas Pharris-Ciurej
U.S. Census Bureau

Sonya R. Porter
U.S. Census Bureau

Jonathan Rothbaum
U.S. Census Bureau

John Voorheis
U.S. Census Bureau

CES 24-23 **April 2024**

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact Sean Wang, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov. To subscribe to the series, please click here.

Abstract

Federal statistical agencies and policymakers have identified a need for integrated systems of household and personal income statistics. This interest marks a recognition that aggregated measures of income, such as GDP or average income growth, tell an incomplete story that may conceal large gaps in well-being between different types of individuals and families. Until recently, longitudinal income data that are rich enough to calculate detailed income statistics and include demographic characteristics, such as race and ethnicity, have not been available. The Mobility, Opportunity, and Volatility Statistics project (MOVS) fills this gap in comprehensive income statistics. Using linked demographic and tax records on the population of U.S. working-age adults, the MOVS project defines households and calculates household income, applying an equivalence scale to create a personal income concept, and then traces the progress of individuals' incomes over time. We then output a set of intermediate statistics by race-ethnicity group, sex, year, base-year state of residence, and base-year income decile. We select the intermediate statistics most useful in developing more complex intragenerational income mobility measures, such as transition matrices, income growth curves, and variance-based volatility statistics. We provide these intermediate statistics as part of a publicly released data tool with downloadable flat files and accompanying documentation. This paper describes the data build process and the output files, including a brief analysis highlighting the structure and content of our main statistics.

[*] Corresponding author: Maggie R. Jones, margaret.r.jones@census.gov. Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project 7511151: CBDRB-FY23-CES014-008, CBDRB-FY23-CES014- 014, CBDRB-FY23-CES014-048). The authors thank Abigail Wozniack, Charles Hokayem, Jim Ziliak, Ken Troske, and participants in NBER's Economics of Mobility conference for helpful feedback. Special thanks to Brian Knop for his expertise in building the Tableau data tool.

1 Introduction

The estimation of statistics on income and earnings growth in the U.S. context has long presented multiple challenges for researchers. One core challenge stems from the lack of access to precise income information at the person or household level due to privacy and confidentiality concerns. While the U.S. is not necessarily unique in prioritizing privacy and confidentiality over social transparency when it comes to income, these priorities stand in contrast to those of some other industrialized countries, where information on income at the person level is more easily attained and occasionally even publicly reported.[1]

Previous research on U.S. income and earnings growth has made some headway through the use of repeated cross-sectional or panel survey data. However, truly understanding mobility as it relates to personal and family well-being requires repeated measures on the same individuals over a long period. Key questions include: What does income growth look like at each point in an initial income distribution? How persistent is a person's year-over-year position in the income distribution? How do mobility patterns vary by demographic group? Until recently, data to answer these questions for the U.S. have been available only as survey data, which suffers from misreporting (Bollinger, 1998), small sample sizes (Browning et al., 2014), and attrition (Kim and Tamborini, 2014). Each of these issues exacerbates the problem of capturing accurate income values for detailed racial and ethnic groups that suffer from small sample sizes in common surveys. In addition, much available survey data provides short time frames, multiple-year gaps, and poor geographic coverage.

The increasing use of administrative records has led to improvements in population coverage and researchers' ability to incorporate repeated measures. In the U.S. context, however, administrative data that capture earnings and income often do not contain information on age, sex, and race-ethnicity. Because these characteristics strongly correlate with income—in terms of levels, inequality, and growth—any analysis of U.S. mobility would be incomplete without taking these characteristics into account.

The Mobility, Opportunity, and Volatility Statistics project (MOVS) uses administrative records and demographic data on a near-complete population of working-age adults to address these challenges. The data-linkage infrastructure at the U.S. Census Bureau allows us to link persons over

[1] Sweden is the best example of this; see, for example, Roine and Waldenström (2008).

data sources and time, a process that captures the population of working-age persons each year. Income and household structure for this population can then be tracked over many years. For each person in the population, our linked microdata contain age, sex, detailed race and Hispanic origin, household structure, and various levels of geography. We then output a publicly available suite of statistics from this population, calculated within cells defined by income decile in the base year, sex, race-ethnicity, year, and geography. The resulting "intermediate" statistics—in that they may stand alone or be used by researchers as the components of more complex mobility measures—disseminate valuable information to the public on patterns of income mobility and volatility. Moreover, MOVS is released as a data tool and sets of flat files, which allow researchers to make their own normative decisions on how to express aggregate income growth, volatility, and mobility; these intermediate statistics will also be valuable in their own right for geographic-level analyses.

In this paper, we present a narrative of the background and goals of the MOVS project, a detailed description of the data build and the development of the intermediate statistics, and an overview of the files released as part of the initial base year of the MOVS project (2005). We expect the MOVS project to release additional data files in the future.

2 Background and related data

Our work adds to recent innovative work from other federal institutions estimating income statistics such as year-to-year volatility and its time trend (Dahl et al., 2007), income inequality (Hungerford, 2011; DeNavas-Walt and Proctor, 2015; Fixler et al., 2020; Kondo et al., 2023), and definitions of income class (Elwell, 2014). It is also related to initiatives at the federal level that provide public-use data on inergenerational mobility (the Opportunity Atlas[2]) and the Bureau of Economic Analysis's (BEA) individual distributional accounts.[3] These new data products constitute a move to create official federal statistics where none have been available previously.

Researchers have long used tax data to examine patterns of income mobility and volatility, although these studies have necessarily been limited in their ability to examine correlates such as race and family structure (Kopczuk et al., 2010; Auten and Splinter, 2022). Recent work has

[2] www.opportunityatlas.org/

[3] www.bea.gov/data/special-topics/distribution-of-personal-income

expanded these analyses in innovative ways by examining pseudo-households (Larrimore et al., 2020) or superimposing demographic information onto administrative aggregates (Piketty et al., 2018). This research represents a considerable improvement over studies that use tax records alone, but may still tell an incomplete story due to the lack of individual-level information on crucial demographic characteristics. Key papers that inform our research in this area include those that use either survey data alone (Bloome and Western, 2011; Bloome, 2014; Van Kerm, 2009) or in combination with administrative records (Akee et al., 2019). The recent ability to link administrative data, typically tax records, with demographic data at the individual level has generated a series of new papers studying these concepts.

Previous work—based on surveys or on administrative records—has focused on three broad themes in income measurement: inequality, mobility, and volatility. These concepts may apply to an individual or household within the course of its existence (*intra*generational); or evaluated between a household or individual and their offspring (*inter*generational). We briefly define these concepts for the sake of clarity, but note that considerable ongoing work exists seeking to precisely and thoroughly define them, the full nuances of which are discussed in other work (see Burkhauser and Couch (2011)).

Inequality: the variation in income between the lower and upper parts of the distribution. Income inequality is a fairly straightforward concept, where some distance metric (such as the ratio of the 90th percentile of the income distribution to the 10th percentile) defines how "equal" or "unequal" a society is. Survey data has long been used to measure inequality (Reardon and Bischoff, 2011; Rose, 2016; Snipp and Cheung, 2016).[4] A particularly robust literature using administrative records alone or in tandem with survey data or aggregates describes income inequality in the U.S. by documenting its increase over time (DeBacker et al., 2013; Piketty et al., 2018) after a period of decrease beginning in the 1940s (Piketty and Saez, 2003; Kopczuk et al., 2010).

Mobility: how often or how far individuals or households move between the various parts of the distribution. Measuring income mobility presents researchers with conceptual challenges. In decisions over what "matters" when we think of income growth patterns, researchers have come to distinguish between shifts in market income that reflect changes in the overall structure of

[4] See also the regularly released Census Income Report: https://www.census.gov/library/publications/2023/demo/p60-279.html

the distribution ("structural mobility") or changes in each individual's position in the distribution ("exchange mobility").[5] Although a swapping of position in a given income distribution can provide important information on how one person might be faring (to another person's detriment), it fails to provide information on absolute growth in income and how growth is shared out among individuals.

Measures of mobility that do not take into account the direction of movement ("upward" or "downward") by different groups may miss patterns important to assessing overall well-being. Although high mobility, which is a low association between origin and destination, is linked to "a more open society and greater equality of opportunity" (Jäntti and Jenkins, 2015), if one group displays significant downward mobility and another displays significant upward mobility of the same magnitude, summarizing this information as though the two groups have equivalent mobility loses important nuance. We center our provision of intermediate statistics, rather than summary measures such as a Shorrocks index, to illustrate how observed changes in mobility have complex implications when considering the direction of mobility for each demographic group.

Volatility: the number and magnitude of income changes, upward and downward, for a given household or individual over some period. The presence of increasing or decreasing volatility in income in the United States has been a source of much debate (Dynan et al., 2012; Hardy and Ziliak, 2014; McKinney and Abowd, 2023; Moffitt and Zhang, 2018; Carr and Wiemers, 2018). Moffitt et al. (2023) explores the role of differing data sources in fostering this debate. Researchers face an additional challenge of standardizing definitions of volatility. The unit of measurement has not been standardized, and neither has the time period over which the income change is calculated. A household may appear volatile within a month, but non-volatile across months, for example. Based on our available data, MOVS provides statistics on volatility from year to year, including information on the prevalence of large swings in income.

3 Input Files

We bring together a variety of data to measure inequality, mobility, and volatility, including the 2000 decennial census; the American Community Survey (ACS); Internal Revenue Service forms

[5] Similarly, volatility may be reported using changes in absolute income or changes in rank within the distribution, both of which are best interpreted in the context of the growth of the overall economy and the various points in the distribution.

1040, W-2, and 1099; files from the Department of Housing and Urban Development (HUD); the Census Bureau's master list of social security numbers, the Numerical Identification Files from the Social Security Administration (the Census Numident); the Master Address File (MAF); and further address history and parent-child linkage files built from survey and administrative records.

The Census Bureau processes input data files via the Person Identification Validation System (PVS) (Wagner and Layne, 2014) in order to place a unique identifier, called a Protected Identification Key (PIK). This identifier is invariant within person over time, allowing us to match all data sources and years at the individual level. PVS takes key variables—social security numbers, names, dates of birth, and so forth—in each dataset and compares them against a master reference file to place the PIK.[6] When address information is available in a dataset, PVS also attempts to place a Master Address File Unit ID (MAFID), which facilitates linking at the address level. Below, we briefly outline each data set and how it is used to create the MOVS data.

The initial building block for our project is the 2000 decennial census. This census is intended to capture all individuals living in the U.S. on April 1, 2000. We use these data to define our initial population of working-age individuals and as our preferred source of race and ethnicity information. To add additional demographic information, we supplement these data with the Numident, which covers all individuals who have received a social security number (SSN). We use the Numident as an administrative source of sex, citizenship, and dates of birth and death. It also aids us in expanding our baseline population to account for population changes between 2000 and 2005. When individuals in our baseline population do not have information in the 2000 Census on their race and ethnicity, we use the Census Best Race and Ethnicity file. This file combines information from multiple surveys and administrative records to assign a race and ethnicity to an individual.[7]

Our main source of income information is individual income tax returns, IRS Form 1040s. We use total money income (pre-tax) as reported by IRS as our main income measure. However, many individuals do not file an individual income tax return but may still have some income. To measure income for non-filers, we turn to wage and salary information returns, also called Form W-2s. Individuals are supposed to receive these forms if they have any wage and salary earnings in a calendar year, regardless of whether they file a tax return.

[6] These key matching variables are stripped from the data before researchers gain access. All analysis proceeds on anonymous data.

[7] For more information on the Census Best Race and Ethnicity file, see Ennis et al. (2018).

Finally, for a small group of individuals who do not file a tax return and do not receive a W-2, we turn to administrative data from HUD. The HUD Public and Indian Housing Information Center and Tenant Rental Assistance Certification Systems (PIC-TRACS) Longitudinal Files contain information on individuals residing in public housing, participating in the Housing Choice Voucher Program, or receiving project-based rental assistance. As an additional measure of income information, we use the income that households report to housing authorities to determine their eligibility for HUD programs.

To measure household composition, our main source of information is also the IRS Form 1040s. The Form 1040 extracts that Census receives from the IRS include information related to filing status, address at the time of filing, and PIKs for the primary filer, secondary filer, and up to four dependents. We use this information to define which individuals live together—either based on whether they live at the same address or appear on a tax return together—and to calculate the number of adults and dependents (usually children and young adults attending school) living in a household.

We supplement the 1040s with geographic and household information from the HUD PIC-TRACS data; the Census Composite Person Record (CPR), Master Address File Auxiliary Reference File (MAF-ARF), and Census Household Composition Key (CHCK); and information returns from the IRS, also called 1099s. The CPR, MAF-ARF, and CHCK are all composite files based on administrative records. The CPR and MAF-ARF primarily contain annual information on individuals' addresses, while the CHCK data links children to their parents.[8] The IRS 1099 data contains the addresses to which various information returns were sent.

4 Data Build

The goals of the MOVS project are, first, to assemble annual, household size-adjusted market income information for the population of working-age adults in the U.S.; and, second, to report out the intermediate statistics that form the components of common mobility measures.[9] What follows are details on how we define the analysis population, construct households, and define our

[8] For more information on the CHCK data see Genadek et al. (2021).

[9] For example, rather than producing transition matrices, we produce origin deciles and destination quintiles, plus underlying counts, so that external researchers can produce their own.

income concept. Figure 1 displays graphically each step in the data build and analysis population selection.

4.1 Building the Analysis Population

We start by defining the analysis population. First, using the 2000 decennial census, we collect all individuals who were born between 1955 and 1980, which gives us adults between the ages of 25 and 50 in 2005 (39 and 64 in 2019). Throughout, we refer to this population as "working-age adults" to reflect our interest in a cohort that is bracketed by labor market entry for the youngest ages at the start of the period and retirement for the oldest ages at the end of the period. Second, we link this population to the Numident and perform several adjustments. For observations that receive a PIK in the 2000 decennial census (about 84 percent of the analysis population), we remove those who die before 2005. From the list of Numident observations in the age range who do not appear in the decennial census, we add individuals who received a social security number between 2001 and 2005 and are either citizens or legal residents authorized to work, thus updating the analysis population to 2005. Third, we link anyone in the age range from the Numident to tax year 1999 1040s filed from overseas, capturing citizens who happened not to be in the U.S. at the time of the 2000 enumeration.

Our choices leave us with a 2005 working-age population where approximately 84 percent of the individuals have a person identifier that allows us to find them in future administrative records (see Table 1). In previous work using similar data, the analysis population was defined as either 1040 filers (the single-generation case in Akee et al. (2019)), *or* the population of children claimed on a 1040 who also appear in the Numident (the intergenerational case in Chetty et al. (2020)). Neither choice is entirely satisfactory, because in both cases, selection into the analysis population relies on adults' 1040 filing behavior. Nonfilers tend to be low-income adults without children. Thus, focusing only on the population of filers could miss important dynamics at the lower end of the income distribution.[10] In basing our sample on the 2000 census data, which is designed to cover all U.S. residents, the MOVS project avoids some of these concerns. However, we still must contend with the fact that some decennial census records are not assigned a PIK, and previous research

[10] The potential for bias is smaller in the case of Chetty et al. (2020), where the analysis sample of children was based on parent claiming. Because of the tax advantages of claiming a child, a large portion of the U.S. child population was reliably captured (Gee et al., 2022).

shows that these non-linkable individuals tend to be more disadvantaged (Bond et al., 2014). To reduce potential bias from only using linkable records, we adopt an inverse-probability-weighting (IPW) strategy, described in more detail in the next subsection.

4.2 Matching the population

Our underlying data consist of working-age individuals in 2005 who have PIKs. To make our population more representative of the entire working-age population within each year, we use inverse probability weights to give greater weight to individuals observed in the data whose characteristics are similar to individuals who do not receive PIKs.

Although our underlying population does not change over time (see Table 2 for summary statistics at baseline), the year-by-year makeup of the true population of working-age adults, as reflected in the ACS, may experience changes. We also may imperfectly match our target population to the true population through the process of PIK placement. This process has undergone improvements over time, increasing the probability of identifier placement. As a result of of these improvements, a person who appears in both the 2000 decennial census and, for example, the 2015 ACS may have a higher probability of receiving a PIK in the 2015 ACS than in the 2000 decennial census. To account for this time-varying slippage between our analysis population and the "true" population we wish to reflect, we retain our original linkable analysis population and, in every year, supplement this unchanging set of observations with working-age individuals in the ACS who do not receive a PIK. These unlinkable persons form the basis, in each year, of an IPW strategy that adjusts our statistics to account for selection into sample (i.e., the probability of receiving a PIK).[11]

In each year, we combine our data with the set of individuals in the relevant year of ACS data who do not receive a PIK and whose year of entry to the U.S. was before 2005.[12] We use the ACS because it contains similar demographic variables to our population data, which allows us to model the probability of receiving a PIK in our population. We assume that the individuals without PIKs in the ACS for a given year are representative of the individuals who would not receive a PIK in our administrative data sources for the same year since the files were likely processed using similar PVS technology.

[11] For an overview of IPW strategies, see Wooldridge (2007).

[12] We wish to capture only people who were in the U.S. in 2005 or earlier so as to avoid including population changes due to new migration.

We use a logit model to estimate the probability of receiving a PIK. Our model includes indicators for household size, number of children in the household, bins of age, state of residence, and interactions of sex, race, and marital status. Our final weights are the inverse of the predicted probability of receiving a PIK in a given year. The observations without a PIK are dropped after contributing their information for the IPW re-weighting.

Using our inverse probability weights, our baseline population looks remarkably similar over time to public-use ACS populations defined using the same age range and restricting to citizens. In Table 3, we report the sex composition and race-ethnicity for our weighted working-age population along with analogous statistics for individuals aged 25–50 in the 2005 ACS and aged 39 and 64 in 2019. The statistics closely match across the two datasets.

4.3 Constructing Households

To develop a household definition of income, we first need to correctly assign addresses to individuals, and then group individuals into households. Most of our data sources contain the address identifier, the MAFID. Only in a small fraction of cases do we rely on a non-address-based household identifier to group households.

4.3.1 Address Identification

We acquire address information both for everyone in the focal population and for anyone with whom they might live. In each year, we begin with the set of all individuals who appear in the Numident and are alive for at least one day during the focal year. We search across our data sources for information on individuals' locations in the year of interest, focusing on address (identified by MAFID), state, and ZIP code of residence. We prioritize location information from IRS 1040s,[13] followed by information from (in order of preference) HUD,[14] IRS information returns (1099s and W-2s), and the CPR (for 2004-2009) or the MAF-ARF (for 2010 onwards).

Finally, if a child does not appear in any other data source and their parent's address comes from the 1099s or CPR or MAF-ARF (i.e., they are neither Form 1040 filers nor receiving HUD

[13] We assume that the address on a Form 1040 applies to all individuals in a tax unit, including primary and secondary filers and up to four dependents. We use the address from the 1040 that was filed during the relevant year.

[14] HUD provides residence information and a family roster that is comparable to what is available on a 1040.

assistance), then we assign the child the address of their parent, favoring the mother's address over the father's address, if they are different.

Once we have constructed preliminary files for each year, we try to fill in missing information using data from other years. If an individual is missing geographic information in a given year, we check if they have geographic information in the year before and year after the focal year. If this geographic information matches in both the year prior and the year after, then we assign the individual that geographic information in the focal year.

Table 4 provides a brief breakdown of MAFID sources in 2005. We find a MAFID in the 1040 data for approximately 73 percent of the analysis population. The next largest contributor of MAFID information is IRS information returns (13 percent of the population). The HUD, CPR, and CHCK data together provide MAFID information for slightly less than 3 percent of individuals.

4.3.2 Grouping Individuals

Once we have collected available location and household information for individuals in our data, we group individuals into households. We use both MAFID and household information from the data sources to construct our households. Household members may be grouped into a MAFID or into a household identifier when a physical location is absent or incomplete.

We apply the following processing rules for determining whether to treat individuals observed in the same MAFID as members of the same household or as separate households. The steps are applied in order. Within each step, we check whether adding more individuals to a household would create a household with more than 10 individuals. If this is the case, we do not add more individuals to the household and do not carry the household forward to later steps. For example, in step 2, if combining a Form 1040 household with a HUD household in the same MAFID would create a household with 11 individuals, we do not group the Form 1040 and HUD household together and also do not allow either of the households to be grouped with individuals with MAFIDs from the 1099 or CPR (MAF-ARF in later years) data in later steps.[15]

1. The Form 1040 data provide information on marriage and the number of children in the household through the filing status and dependent claiming fields. Via the secondary filer

[15] Note that, when discussing addresses, W-2 address information is pooled with 1099 information.

field we can assign spouses, and through the dependent fields we can confirm the identity of the first four dependents (fields regarding the number of children claimed gives us information for calculating the equivalence scale). We assume that a unique 1040-MAFID pair is a complete household.

2. When we see multiple tax units in a single MAFID in the Form 1040 data, we group these tax units into one household. Tax filing generally occurs between February and April of a single year, so we assume that individuals with 1040 forms listing the same MAFID are co-residing. This may include multiple families living together, or it may reflect children or older adults who are dependent on another 1040 filer in the household but who file their own 1040.

3. When we see individuals in a MAFID in the Form 1040 data and individuals in the same MAFID in the HUD data for the same year, we do the following: If at least one person in the Form 1040 tax unit appears in the same HUD assistance unit as a person only observed in HUD, then we group them together. Otherwise, we create separate units, as the HUD household may have lived in the MAFID at a different time during the year from the 1040 household.

4. When we see individuals in a Form 1040 MAFID and individuals in the same MAFID in the 1099, CPR, MAF-ARF, or CHCK data in the same year, we group the 1099/CPR/MAF-ARF/CHCK individuals with the Form 1040 household if the 1099/CPR/MAF-ARF/CHCK individuals appear in the same MAFID as the 1040 primary filer(s) in either the year preceding or following the relevant year. Note that the MAFID in the preceding or following year need not be the same MAFID as the current year. In other words, we only group Form 1040 and 1099/CPR/MAF-ARF/CHCK individuals together if they appear to reside together in more than one year. We impose this rule to account for families moving in the middle of a tax year: we want to avoid grouping together individuals who lived at the same address at different times during the same year.

5. Next, in line with Larrimore et al. (2021), we use MAFID to collect into households those receiving a Form W-2 or 1099 at the same address in the same year, or appearing in the CPR, MAF-ARF, or CHCK in the same year. In the absence of additional household information

about these individuals, we default toward assuming co-residence and a sharing of resources. We limit the size of these households to 10 following Larrimore et al. (2021).

6. When we see multiple households with the same MAFID in the HUD data, we keep these households separate. HUD household units are designed to include all individuals residing together. Thus, we assume that groups appearing as separate households in the HUD data do not reside together, but rather lived in the MAFID at two different times during the same year.

7. When we see individuals in a MAFID in the 1099, CPR, MAF-ARF, or CHCK data and individuals in the same MAFID from HUD, we group them separately. As with the multiple HUD household rule above, we default towards assuming that these households are more likely to live in the MAFID at different times during the year rather than being part of the same household. Unlike Form 1040 filing, we cannot place 1099/CPR/MAF-ARF/CHCK recipients at a location at a specific time in the year. It also seems unlikely that individuals living in a HUD household would not be included when the household was certified. Thus we do not include individuals whose 1099/CPR/MAF-ARF/CHCK MAFID matches a HUD household MAFID as part of the HUD unit unless they are specifically listed in the unit.

8. When individuals appear in the Form 1040, HUD, or CPR data but do not have a MAFID, we use their unit identifiers to group them into households. All individuals appearing in the same tax unit are included in the same household, and all individuals appearing in the same HUD household ID are grouped into the same household. If a CPR household ID (HUID) contains ten individuals or fewer, then we group the individuals into the same household. Otherwise, we keep the CPR individuals as singleton households.

Even in the absence of a MAFID, we are able to assign a key grouping variable—state in 2005—to close to 100 percent of observations through state or ZIP Code being available on the source data files.

4.4 Capturing Equivalized Income

For our individual measure of income, we use an equivalized concept (Buhmann et al., 1988), where income is summed over a household, and each working-age household member is allotted total income divided by the square root of the number of individuals in the household. Such a calculation requires a measure of total income and a measure of household structure over time for each member of our analysis population. We index future dollars to 2005 dollars using the Bureau of Labor Statistics' Consumer Price Index for all Urban Consumers (CPI-U).

4.4.1 Creating Total Household Income

Each year's income values come from Form 1040 filings, W-2 reports, and income reported to HUD. Our primary measure of income is total money income, which includes most sources of income from the "income" section of a 1040 (Meyer et al., 2020), including wages, self-employment income, ordinary dividends, social security benefits, and rental income.[16] Importantly, it does not include capital gains or losses. The definition thus aligns closely with the Census Bureau's definition of money income[17] that is used in the agency's official reports. We construct household income by summing total money income for all tax units in the household. If an individual is a non-filer, we instead add the wages reported on their W-2s to the incomes of other household members. If no one in a household files a 1040 or receives a W-2, then we define household income as the sum of all individual income reported to HUD. This last definition accounts for less than 1 percent of cases.

We find household income for approximately 90 percent of our working-age population in 2005. For the core set of intermediate statistics, we require persons to have an income report in the base year. However, in a supplementary file, we provide statistics for those whose base-year income is missing. Each year's income value, including the base year, is winsorized at the bottom and top of the distribution by a percentage that brings the lowest possible value to greater than zero. In later years, income values less than or equal to one and missing values are assigned one dollar of income

[16] The total money income variable that IRS provides to Census generally includes the gross pension income individuals report on 1040s. However, in 2018, gross pensions are not included. We use IRS 1099-R information returns to impute gross pension income in 2018, taking into account whether individuals appeared to report their gross pensions in 2017 or 2019. The 1099-R extracts that Census receives do not include all types of distributions that would normally be reported in gross pensions, including direct rollovers, Section 1035 exchanges, and Roth conversions (Bee and Mitchell, 2017). Thus, the levels of total money income in 2018 still show a dip for some groups.

[17] https://www.census.gov/topics/income-poverty/income/about/glossary/alternative-measures.html

for that year. We take this step so that researchers may easily calculate mobility measures that rely on the natural log of income, described in more detail in the next section. In a supplementary file, we provide suitable statistics, such as mean income within cell and income shares, using unwinsorized income.

5 Intragenerational Intermediate Statistics

The final step in the process involves calculating the suite of statistics for public consumption. A particular challenge in producing statistics on intra- or intergenerational mobility involves the reporting of income. Observations in our data may report positive, zero, or negative income—or they may not report at all for either the base year or some other year. Meanwhile, many of the key statistics for describing income growth, deviations from income, or relative mobility rely on the natural log of income, which is not defined when incomes are non-positive. Solutions such as applying a zero to missings or only reporting key statistics for the subset of those with strictly positive income are problematic when attempting to “say something” about income patterns for the full population.

In the case of 1040 income, an additional challenge arises when we think about how to handle income earners who are older than 18 but less than 25, as a large portion of these will be students and still dependent on parental support. Restricting our analysis population only to those over 25 seems inadequate for painting the full picture of income growth in the U.S.

MOVS approaches issues of income intransigence and possible dependency by releasing a set of five files, each of which we describe in detail. The main file—i.e., the winsorized file—uses the winsorized value of income as described in subsection 4.4.1; the definition of income used for it is intended to provide a tractable file of intermediate statistics where individuals represent the overwhelming majority of the U.S. working-age population. Other files round out the story of income growth. Further data releases will also include metro-level files in addition to the state-level files.

5.1 Main File: Winsorized Income

We define a set of cells based on sex, race-ethnicity, 2005 state of residence, and decile of 2005 age-adjusted and equivalized income rank. To create the age-adjusted deciles, we rank all individuals born in a specific calendar year by their 2005 equivalized income and assign their percentile position based on this ranking. Thus, each decile contains roughly the same number of individuals from each birth cohort. Without age adjustments, the higher income deciles would be dominated by individuals further into their career progression while individuals early in their career would tend to fall into the lowest income deciles. The calculation of decile position and the cell-level statistics use the 2005 inverse probability weights to adjust for sample selection.

Within each cell, we provide the statistics listed in Table 5. Mean income and income changes from both the base year and year-to-year provide broad descriptive statistics on how income changed from 2005 to 2019. We provide these statistics for both equivalized income and total household income (no equivalence scale applied). Our preferred income change variable is the mean arc percent change in equivalized income. The arc percent change is defined as the difference in income between two years divided by the average of the income for those same two years. We calculate this amount at the individual level and then average across individuals within a group. We prefer this measure over the percent change—the difference in income from one year to the next, divided by the income in the first year—because the arc percent change can handle cases where income in the first year is equal to zero. If individuals have income equal to zero in the first (second) year, their arc percent change is equal to two (negative two). If income is equal to zero in both years, we set the arc percent change to zero. We also prefer the mean arc percent change over the mean change in log income since the concavity of the log function gives disproportionate weight to income losses, resulting in negative means even when income gains exceeded income losses for a group. While the arc percent change function is also somewhat concave, its boundedness limits the extent to which losses can be over-weighted. Nevertheless, we also provide statistics using log income as this is a widely used measure in the income mobility literature, and we simply caution users from relying only on the log income variables.

Because equivalized income within a cell may change due to movements of both income and family structure, we provide variables on base-year-to-year and year-to-year changes in the number of total family members, number of adults, and number of children, as well as the rates of year-over-year changes in tax filing status that would indicate marriage (or remarriage) and divorce. We also provide two measures that allow for separability of changes in log equivalized income into two parts: that due to changes in log unequivalized income and that due to changes in the log of the scale (the square root of household size).

Beyond changes in income amounts, the statistics we provide shed light on how individuals change their position in the income distribution over time. For each year, we calculate the probability that individuals have income placing them in the five income quintiles for the year. As with the base year deciles, these income quintiles are age-adjusted so that each birth-year cohort has its own income quintile definitions.

We include three measures of income volatility. We provide the percent of individuals who experience an increase in income greater than 25 percent of their prior-year income and the percent of individuals who experience a decrease in income greater than 25 percent.[18] These two statistics succinctly describe the percent of individuals who experience large income swings from year to year. The final volatility statistic is the variance of the arc percent change in income. This is a widely used statistic in the economics literature.[19] Relative to the mean arc percent change variable discussed earlier, the variance of the arc percent change puts greater weight on large swings in income.

Finally, we note that individuals are categorized into states based on their state of residence in 2005. Even if individuals move to another state after 2005, they are still included in the cell for their 2005 state. We include two measures to describe how often moves across state lines occur: (1) the probability that an individual's state of residence in one year is different from their state of residence in the previous year and (2) the probability that an individual's state of residence in one year is different than their state of residence in the base year.

In each cell, we additionally provide the IPW-adjusted counts of observations. We used the number of unique tax forms within a cell to determine which cells had too few observations to meet IRS and Census disclosure thresholds. The IPW-adjusted counts are used to develop the upper-level aggregates described in section 5, and they may be used by researchers to correctly weight statistics when combining information across cells.

[18] Individuals with zero income in the prior year and positive income in the next year are counted as having an increase in income greater than 25 percent.

[19] See for example, Moffitt et al. (2023).

5.2 Supplemental Files

We briefly outline the supplemental files we plan to release at a later date. Relative to the main file, these additional files will focus on different measures of income and income mobility and different baseline populations.

5.2.1 Unwinsorized data

Negative total money income or AGI reported on a 1040 indicates a taxpayer, possibly with high permanent income, who experiences business losses or negative capital gains. By winsorizing for negative or zero income, we also trim a portion of the highest-income taxpayers, many of whom may move between reporting income less than zero and very high income. Thus we provide a file of unwinsorized information, where we retain zero and negative total money income values and do not winsorize top incomes. Because of the importance of unwinsorized values at the top of the distribution in the determination of income shares, we reserve the release of income share statistics to this file.

5.2.2 Young adult file

Assessing the condition of younger workers in our full file was complicated by student status, which allows 18- to 24-year-olds to be claimed by their parents. Our young adult file will collect those aged 18–24 and track their income over time, while taking into account their dependency status.

5.2.3 Missing income file

For each base year we produce, we plan to restrict observations to those who are connected to reported income in the base year. For 2005, this means that approximately 8 million working-age adults are dropped from the analysis population (about 5.8 percent). However, many of these observations may be found in later years of income data. This supplementary file will provide statistics on mobility for this group when starting income is missing.

5.2.4 Persistence of low income file

Using relative measures (based on median income) and an absolute measure that aligns as closely as possible with the U.S. official poverty measure, we will provide statistics on the depth and persistence of low income for working age adults.

6 Summary Results from the Main File

In this section we provide a brief overview of the main file and a key summary result that highlights the file's contents and structure.

Table 6 reports on the levels of aggregation in the main file and the number of observations in each level of aggregation. For the most granular level of aggregation, there are 91,800 observations, accounting for 10 deciles by 2 sexes by 6 race-ethnicity groups by 51 states by 15 years. The upper aggregates are provided mainly for convenience—we simply took the finer-level cell statistics and their associated IPW-weighted counts to correctly calculate cell contents for combined categories.

There are cases in which a cell does not have enough coverage in every year to meet disclosure thresholds, in which case these cells appear as missing. However, in all but one case of suppression there was enough support to report information at the decile-by-race-by-state-by-year level. In other words, we could report the information for the two sexes combined. One decile-by-race-by-state-by-year cell also fell below the disclosure thresholds, so we suppressed another decile-by-race-by-state cell within the same decile and state.[20]

In Figure 2, we show an example of using the intermediate statistics from the main file in a simple analysis. Here, we choose the aggregation level for decile, sex, and race-ethnicity (i.e., we do not look separately by state) to trace out mean incomes for men whose base-year income put them in the 5th decile (i.e., between the 40th and 50th percentile). By construction, mean income is similar across race-ethnicity groups in the base year of 2005. Thereafter, average incomes quickly diverge, with White and Asian non-Hispanic men experiencing much more robust income growth compared with other groups. While all groups experience a stagnation or drop in average income around 2009 in tandem with the Great Recession, White and Asian non-Hispanic men also

[20] Specifically, we suppress statistics for both non-Hispanic AIAN and non-Hispanic Other Race individuals in the second decile in the District of Columbia.

recover more quickly than other groups and continue on a stronger upward trajectory. Figure 3 shows the same analysis for women. Patterns are generally similar, although the Hispanic and Other non-Hispanic groups appear to experience slightly better income growth beginning around 2015–2016.

Recall that our population is first observed when everyone is of prime working age: 25 to 50. The trajectory observed in the graphs thus displays an upward pattern in real average incomes due to expected income growth for working-age persons over time. The dip seen for all groups beginning in 2009, and lasting for some groups as late as 2012, is a break in expected upward growth in income resulting from the recession. However, we should note that these growth patterns would look different if we "refreshed" the population using observations who age into working age. Further years of data (2006 onward) will provide researchers with refreshed samples of working-age adults, whose income distributions and trajectories may be compared with the 2005 results.

7 Conclusion

The Mobility, Opportunity, and Volatility Statistics project (MOVS) uses administrative records and population-level data on age, sex, race-ethnicity, and base-year state of residence to address the lack of regularly released statistics on income mobility in the U.S. The data-linkage infrastructure at the U.S. Census Bureau allows us to link persons to their income and household data over time, leading to annual populations of working-age persons who can be tracked for multiple years.

This paper is a narrative of the data build, providing an overview of the source files and the decision points used to generate the MOVS microdata. We also describe how our choice of released statistics relates to previous mobility literature and is meant for both layperson use and researchers' more complex mobility analyses.

We publicly release suites of intermediate statistics, calculated within cells defined by income decile in each base year, sex, race-ethnicity, and geography, that will inform the public on income growth by group and allow researchers to calculate their own final mobility measures. Released as a data tool accompanied by sets of flat files, MOVS allows researchers to make their own normative decisions on how to express income growth, volatility, and mobility; these intermediate statistics will also be useful in their own right for state- or, in future, metro-level analyses.

A look at a single intermediate statistic—average incomes from 2005 to 2019 of men and women of different race-ethnicity groups who were in the middle of the income distribution in 2005—provides key information on differential patterns of income growth for these groups. We find that White and Asian non-Hispanic men and women experienced considerably greater income growth over our 15-year period than did members of other groups.

This paper covers the first base year of data release (2005) and provides details on the main file of statistics. We hope the research community will use these data to uncover additional mobility patterns and incorporate these statistics as part of their larger research agenda.

References

Akee, R., M. R. Jones, and S. R. Porter (2019). Race matters: Income shares, income inequality, and income mobility for all U.S. races. *Demography 56*(3), 999–1021.

Auten, G. and D. Splinter (2022). Income inequality in the United States: Using tax data to measure long-term trends. Technical report, Washington, DC: Joint Committee on Taxation.

Bee, A. and J. Mitchell (2017). Do Older Americans Have More Income Than We Think? *SESHD Working Paper 2017-39*.

Bloome, D. (2014). Racial inequality trends and the intergenerational persistence of income and family structure. *American Sociological Review 79*(6), 1196–1225.

Bloome, D. and B. Western (2011). Cohort change and racial differences in educational and income mobility. *Social Forces 90*(2), 375–395.

Bollinger, C. R. (1998). Measurement error in the current population survey: A nonparametric look. *Journal of Labor Economics 16*(3), 576–594.

Bond, B., D. J. Brown, A. Luque, and A. O'Hara (2014). The Nature of the Bias When Studying Only Linkable Person Records: Evidence from the American Community Survey. *Center for Administrative Records Research and Applications Working Paper 8*.

Browning, M., T. F. Crossley, and J. Winter (2014). The measurement of household consumption expenditures. *Annu. Rev. Econ. 6*(1), 475–501.

Buhmann, B., L. Rainwater, G. Schmaus, and T. M. Smeeding (1988). Equivalence scales, well-being, inequality, and poverty: sensitivity estimates across ten countries using the luxembourg income study (lis) database. *Review of income and wealth 34*(2), 115–142.

Burkhauser, R. V. and K. A. Couch (2011). Intragenerational Inequality and Intertemporal Mobility. In *The Oxford Handbook of Economic Inequality*, pp. 522–546. Oxford University Press.

Carr, M. D. and E. E. Wiemers (2018). New evidence on earnings volatility in survey and administrative data. In *AEA Papers and Proceedings*, Volume 108, pp. 287–91.

Chetty, R., N. Hendren, M. R. Jones, and S. R. Porter (2020). Race and economic opportunity in the United States: An intergenerational perspective. *The Quarterly Journal of Economics 135*(2), 711–783.

Dahl, M., T. DeLeire, and J. Schwabish (2007). Trends in earnings variability over the past 20 years. Technical report, U.S. Congress, Congressional Budget Office.

DeBacker, J., B. Heim, V. Panousi, S. Ramnath, and I. Vidangos (2013). Rising inequality: transitory or persistent? New evidence from a panel of U.S. tax returns. *Brookings Papers on Economic Activity 2013*(1), 67–142.

DeNavas-Walt, C. and B. D. Proctor (2015). Income and poverty in the United States: 2014. Technical report, U.S. Department of Commerce, U.S. Census Bureau.

Dynan, K., D. Elmendorf, and D. Sichel (2012). The evolution of household income volatility. *The BE Journal of Economic Analysis & Policy 12*(2).

Elwell, C. K. (2014). The Distribution of Household Income and the Middle Class. Technical report, U.S. Congress, Congressional Research Service.

Ennis, S. R., S. R. Porter, J. M. Noon, and E. Zapata (2018). When race and Hispanic origin reporting are discrepant across administrative records and third party sources. *Statistical Journal of the IAOS 34*, 179–189.

Fixler, D. J., M. Gindelsky, and D. S. Johnson (2020). Measuring inequality in the national accounts. Technical report, U.S. Department of Commerce, Bureau of Economic Analysis.

Gee, G., J. Goldin, J. Hancuch, L. Ithai, and V. Vohra (2022). The Claiming of Children on U.S. Tax Returns, 2017-2019. *Available at SSRN 4066787.*

Genadek, K., J. Sanders, and A. J. Stevenson (2021). Measuring U.S. Fertility using Administrative Data from the Census Bureau. *Associate Director for Economic Programs Working Paper Series ADEP-WP-2021-02.*

Hardy, B. and J. P. Ziliak (2014). Decomposing trends in income volatility: The "wild ride" at the top and bottom. *Economic Inquiry 52*(1), 459–476.

Hungerford, T. L. (2011). Changes in the distribution of income among tax filers between 1996 and 2006: The role of labor income, capital income, and tax policy. Technical report, U.S. Congress, Congressional Research Service.

Jäntti, M. and S. P. Jenkins (2015). Income mobility. In *Handbook of income distribution*, Volume 2, pp. 807–935. Elsevier.

Kim, C. and C. R. Tamborini (2014). Response error in earnings: An analysis of the survey of income and program participation matched with administrative data. *Sociological Methods & Research 43*(1), 39–72.

Kondo, I., K. Rinz, N. Gubbay, B. Hawkins, A. Wozniak, and J. Voorheis (2023). Granular income inequality and mobility using IDDA: Exploring patterns across race and ethnicity. Technical report.

Kopczuk, W., E. Saez, and J. Song (2010). Earnings inequality and mobility in the United States: evidence from social security data since 1937. *The Quarterly Journal of Economics 125*(1), 91–128.

Larrimore, J., J. Mortenson, and D. Splinter (2020). Presence and persistence of poverty in U.S. tax data. Technical report, National Bureau of Economic Research.

Larrimore, J., J. Mortenson, and D. Splinter (2021). Household incomes in tax data using addresses to move from tax-unit to household income distributions. *Journal of Human Resources 56*(2), 600–631.

McKinney, K. L. and J. M. Abowd (2023). Male Earnings Volatility in LEHD before, during, and after the Great Recession. *Journal of Business & Economic Statistics 41*(1), 33–39.

Meyer, B. D., D. Wu, G. Finley, P. Langetieg, C. Medalia, M. Payne, and A. Plumley (2020). The accuracy of tax imputations: Estimating tax liabilities and credits using linked survey and administrative data. Technical report, National Bureau of Economic Research.

Moffitt, R., J. Abowd, C. Bollinger, M. Carr, C. Hokayem, K. McKinney, E. Wiemers, S. Zhang, and J. Ziliak (2023). Reconciling Trends in U.S. Male Earnings Volatility: Results from Survey and Administrative Data. *Journal of Business & Economic Statistics*, 1–11.

Moffitt, R. and S. Zhang (2018). Income volatility and the PSID: Past research and new results. In *AEA Papers and Proceedings*, Volume 108, pp. 277–80.

Piketty, T. and E. Saez (2003). Income inequality in the United States, 1913–1998. *The Quarterly Journal of Economics 118*(1), 1–41.

Piketty, T., E. Saez, and G. Zucman (2018). Distributional national accounts: methods and estimates for the United States. *The Quarterly Journal of Economics 133*(2), 553–609.

Reardon, S. F. and K. Bischoff (2011). Income inequality and income segregation. *American Journal of Sociology 116*(4), 1092–1153.

Roine, J. and D. Waldenström (2008). The evolution of top incomes in an egalitarian society: Sweden, 1903–2004. *Journal of Public Economics 92*(1-2), 366–387.

Rose, S. (2016). The growing size and incomes of the upper middle class. Technical report, Urban Institute, Income and Policy Benefits Center.

Snipp, C. M. and S. Y. Cheung (2016). Changes in racial and gender inequality since 1970. *The Annals of the American Academy of Political and Social Science 663*(1), 80–98.

Van Kerm, P. (2009). Income mobility profiles. *Economics Letters 102*(2), 93–95.

Wagner, D. and M. Layne (2014). The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications. *Center for Administrative Records Research and Applications Working paper*.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics 141*(2), 1281–1301.

Table 1: Linkage rates for baseline population

Source	Count	PIKed percent	Final baseline count
Decennial census	108,900,000	83.86%	91,300,000
Numident post 2000	4,017,000	100%	4,017,000
Foreign address 1040	124,000	100%	124,000
		Total	95,441,000

Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY23-CES014-008.

Table 2: Summary statistics at baseline

Variable	Mean	SD
Single person in household	0.104	0.323
Age in 2005	38.03	7.895
Male	0.487	0.529
Household size	3.659	1.948
	Percent	SD
Race-ethnicity		
Non-Hispainc White	65.43	50.3
Non-Hispanic Black	11.35	33.6
Non-Hispanic AIAN	0.71	8.9
Non-Hispanic Asian	4.88	22.3
Non-Hispanic Other	2.14	15.3
Hispanic	15.50	38.3

Source: Decennial census 2000, Numident, Forms 1040 and W-2, HUD, Composite Person Record, MAFARF. Races are defined as non-Hispanic (NH), while Hispanic includes those of any race. DRB approval number: CBDRB-FY24-0197.

Table 3: Mean characteristics of individuals in our data versus the 2005 and 2019 American Community Survey

	2005		2019	
Characteristic	MOVS	ACS	MOVS	ACS
Male	48.68	49.45	48.36	48.96
Race-ethnicity				
Non-Hispanic White	65.43	65.49	66.78	65.31
Non-Hispanic Black	11.35	11.82	11.52	12.25
Non-Hispanic AIAN	0.71	0.73	0.71	0.70
Non-Hispanic Asian	4.88	5.11	4.81	4.91
Non-Hispanic Other	2.14	1.43	2.20	1.81
Hispanic Any Race	15.50	15.42	14.05	15.02

Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK; ACS 2005 PUMS. The Hispanic group includes individuals of any race. The Other Race/Ethnicity group includes individuals who are non-Hispanic Native Hawaiian or Other Pacific Islander, non-Hispanic other race, non-Hispanic more than one race, or non-Hispanic missing race in addition to individuals missing ethnicity information. Columns 2 and 4 show estimates from the public use American Community Survey. DRB approval number: CBDRB-FY23-CES014-048.

Table 4: Source of household information at baseline

Source	Count	Percent of total
MAFID from Form 1040	69,970,000	72.59
MAFID from HUD	854,000	0.95
MAFID from information returns	11,660,000	12.93
MAFID from other source	1,498,000	1.66
Other household information	6,007,000	6.61
Insufficient or no household info	191,000	0.21
Source of income information at baseline		
Form 1040	79,540,000	88.20
Form W-2	5,170,000	5.73
HUD	654,000	0.73
Missing Income	4,816,000	5.34

Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK. DRB approval number: CBDRB-FY24-0197

Table 5: Main File: Variable Names and Definitions

Variable	Definition
Group Variables	
baseyear	Base Year: 2005
decile	Age-adjusted income decile, 2005*
sex	Sex: Male or Female*
raceeth	Race and ethnicity: Hispanic, NH White, NH Black, NH AIAN, NH Asian, or Other*
state	State of residence in 2005*
agglevel	Aggregation Level
Time Variable	
year	Year*
Time-invariant Variables	
fips	State FIPS Code
state_name	State Name
state_abbr	State Name Abbreviation
mageb	Mean Age in Base Year
Time-variant Variables, Income Statistics	
meqinc	Mean Equivalized Income
mleqinc	Mean Log Equivalized Income
muneqinc	Mean Unequivalized Income
pqn1	Percent in First Quintile
pqn2	Percent in Second Quintile
pqn3	Percent in Third Quintile
pqn4	Percent in Fourth Quintile
pqn5	Percent in Fifth Quintile
mdiffleqinc	Mean 1-Year Difference in Log Equivalized Income
mdiffleqincb	Mean Difference in Log Equivalized Income from Base Year
mdiffluneqinc	Mean 1-Year Difference in Log Unequivalized Income
mdiffleqscale	Mean 1-Year Difference in Log Equivalence Scale
pdiffinc_gt25	Percent with 1-Year Income Growth Greater Than 25%
pdiffinc_lt25	Percent with 1-Year Income Loss Greater Than 25%
marcpdiffeqinc	Mean 1-Year Arc Percent Change in Equivalized Income
varcpdiffeqinc	Variance of 1-Year Arc Percent Change in Equivalized Income
marcpdiffeqincb	Mean of Arc Percent Change in Equivalized Income from Base Year
mdiffeqinc	Mean 1-Year Difference in Equivalized Income
mdiffuneqinc	Mean 1-Year Difference in Unequivalized Income
mdiffeqincb	Mean Difference in Equivalized Income from Base Year
mdiffuneqincb	Mean Difference in Unequivalized Income from Base Year
Time-variant Variables, Person and Family Statistics	
count	IPW-Adjusted Count
mfn	Mean Number of Persons in Household
mad	Mean Number of Adults in Household
mkids	Mean Number of Children in Household
pmar	Percent Married
mdiffkids	Mean 1-Year Change in Number of Children in Household
mdiffad	Mean 1-YearChange in Number of Adults in Household
mdifffn	Mean 1-Year Change in Number of Family Members in Household
pdiffstate	Percent with 1-Year Change in State of Residence
pdiffstateb	Percent with Change in State of Residence from Base Year
pdiffdiv	Percent with 1-Year Change from Joint Filing to Single
pdiffmar	Percent with 1-Year Change from Single Filing to Joint

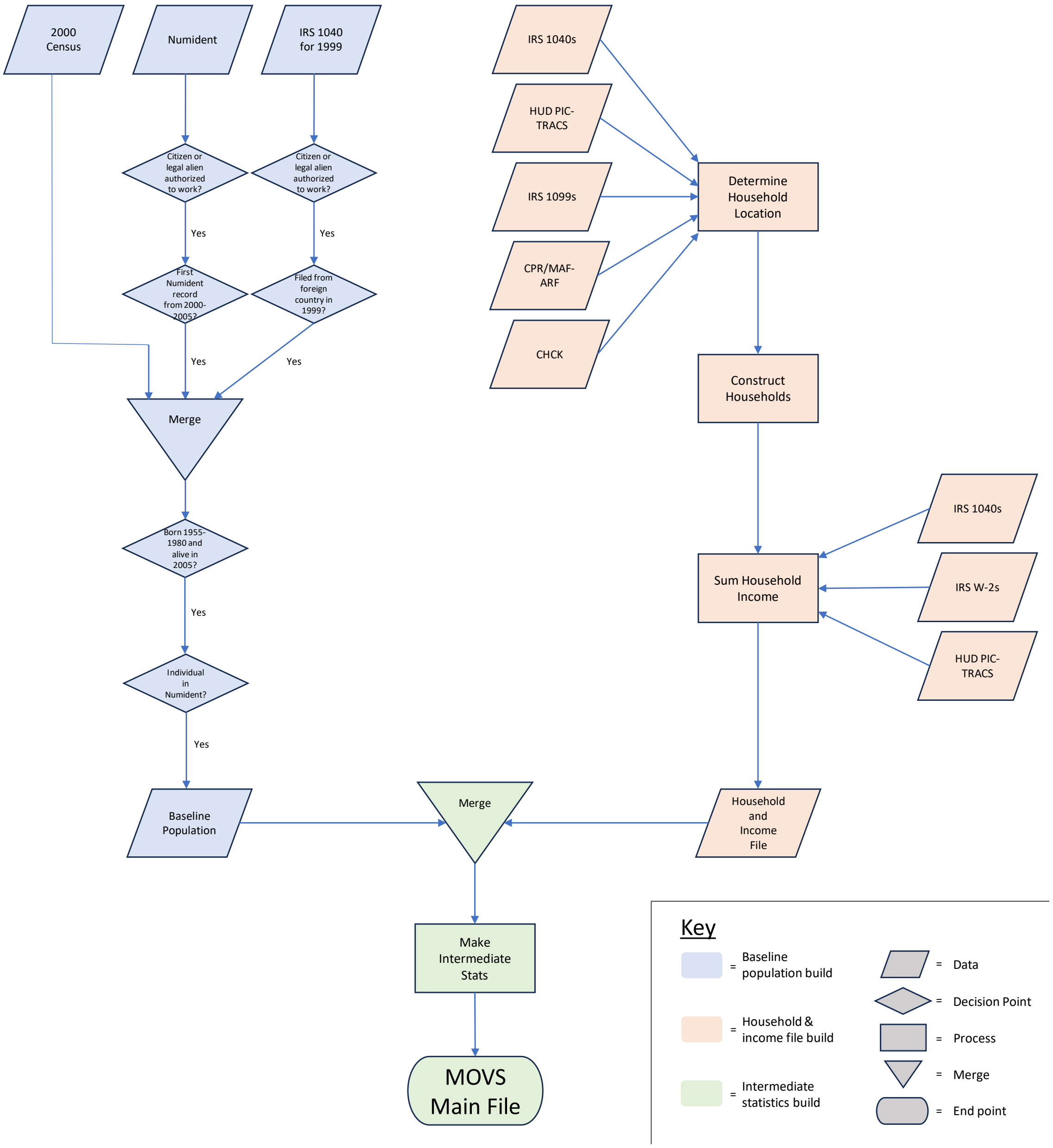
A * indicates a variable that defines cells. Cells are unique in decile, sex, raceeth, state, and year.

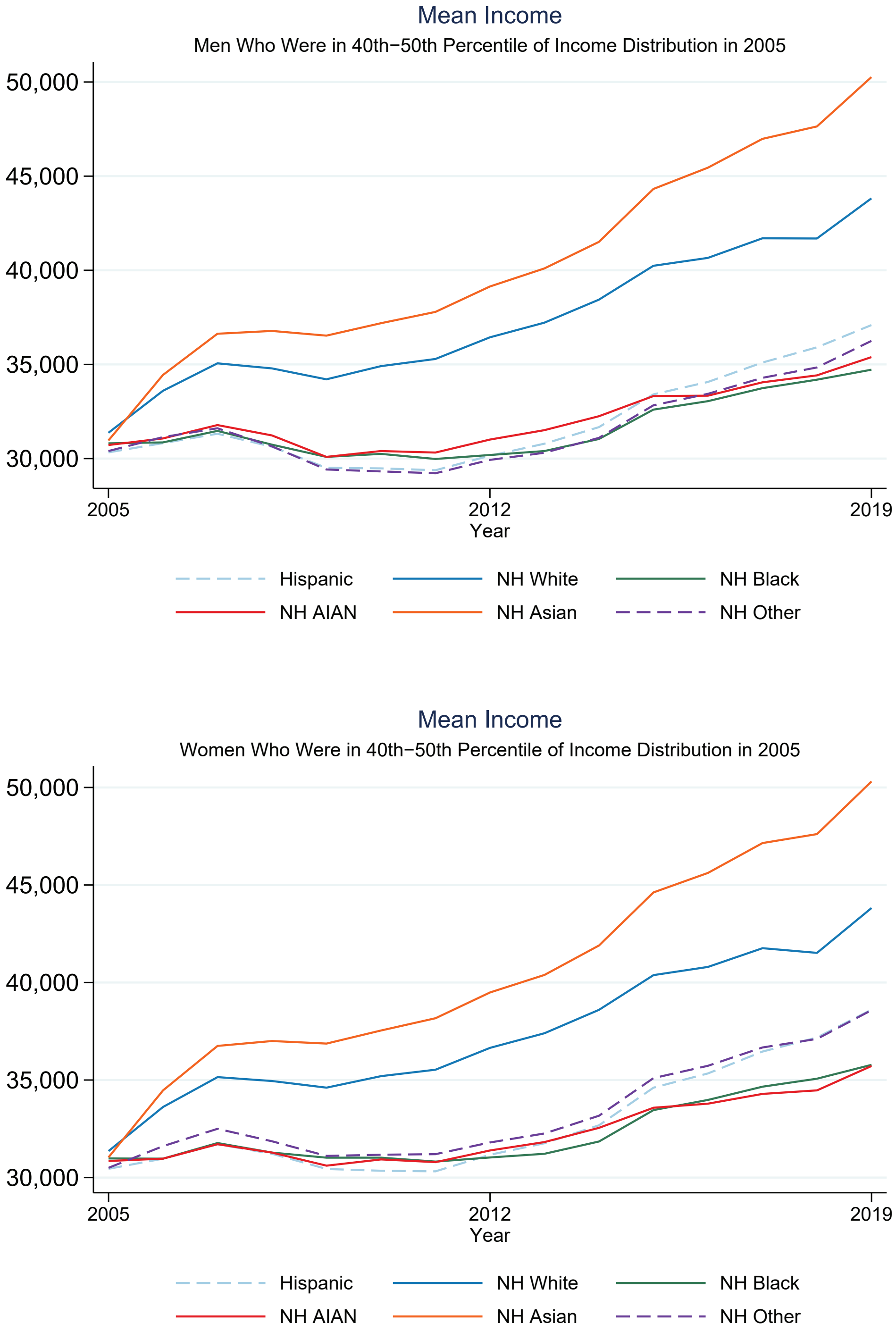
Table 6: Main file: aggregation levels and number of observations

Aggregation level	Total records
National	15
Decile	150
Sex	30
Race-Ethnicity	90
State	765
Decile and Sex	300
Decile and Race-Ethnicity	900
Decile and State	7,650
Sex and Race-Ethnicity	180
Sex and State	1,530
Race-Ethnicity and State	4,590
Decile, Sex, and Race-Ethnicity	1,800
Decile, Sex, and State	15,300
Decile, Race-Ethnicity, and State	45,900
Sex, Race-Ethnicity, and State	9,180
Decile, Sex, Race-Ethnicity, and State	91,800

Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK; ACS 2005 PUMS. The Hispanic group includes individuals of any race. The Other Race/Ethnicity group includes individuals who are non-Hispanic Native Hawaiian or Other Pacific Islander, non-Hispanic other race, non-Hispanic more than one race, or non-Hispanic missing race in addition to individuals missing ethnicity information. Columns 2 and 4 show estimates from the public use American Community Survey. DRB approval number: CBDRB-FY23-CES014-048.

Figure 1: **Outline of Data Build Process**


2000 Census
Numident
IRS 1040 for 1999
Citizen or legal alien authorized to work?
Citizen or legal alien authorized to work?
Yes
Yes
First Numident record from 2000-2005?
Filed from foreign country in 1999?
Yes
Yes
Merge
Born 1955-1980 and alive in 2005?
Yes
Individual in Numident?
Yes
Baseline Population
IRS 1040s
HUD PIC-TRACS
IRS 1099s
CPR/MAF-ARF
CHCK
Determine Household Location
Construct Households
IRS 1040s
IRS W-2s
HUD PIC-TRACS
Sum Household Income
Household and Income File
Merge
Make Intermediate Stats
MOVS Main File
Key
= Baseline population build
= Household & income file build
= Intermediate statistics build
= Data
= Decision Point
= Process
= Merge
= End point



Figures 2 and 3: Mean incomes by race-ethnicity over time for men and women in the middle of the 2005 income distribution. Source: Decennial census 2000; Numident; IRS Forms 1040, W-2, and 1099; HUD PIK-TRACS; Composite Person Record; MAFARF; Census CHCK; ACS 2005 PUMS. DRB approval number: CBDRB-FY24-0197.