ADEP WORKING PAPER SERIES

# Newly Available Individual-Level U.S. Tax Data from 1969-1994

**J. Trent Alexander**
University of Michigan

**David Bleckley**
University of Michigan

**Jonathan Fisher**
Washington Center for Equitable Growth

**Katie Genadek**
U.S. Census Bureau

**Susan Hautaniemi Leonard**
University of Michigan

**Aristotle Magganas**
University of California - Los Angeles

Working Paper 2024-02
February 2024

Associate Directorate for Economic Programs
U.S. Census Bureau
Washington DC 20233

*Disclaimer:* Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. Results were approved for release by the Census Bureau's Disclosure Review Board (Data Management System number: P-7506192), Disclosure Review Board (DRB) approval numbers: CBDRB-FY24-ADEP001-001, CBDRB-FY24-ADEP001-006, CBDRB-FY24-ADEP001-007, and CBDRB-FY22-ERD002-016.

# **Newly Available Individual-Level U.S. Tax Data from 1969-1994**

J. Trent Alexander, University of Michigan
David Bleckley, University of Michigan
Jonathan Fisher, Washington Center for Equitable Growth
*Katie Genadek, U.S. Census Bureau
Susan Hautaniemi Leonard, University of Michigan
Aristotle Magganas, University of California - Los Angeles

*contact author with questions: katie.r.genadek@census.gov

ADEP Working Paper 2024-02
February 2024

**Abstract:** This paper describes a series of linkable individual-level data from late-twentieth century federal income tax returns. The full universe Internal Revenue Service (IRS) Form 1040 files from 1969, 1974, 1979, 1984, 1989, and 1994 were held at the Census Bureau since originally being delivered by the IRS shortly after each year's tax returns were processed. The data were recently made usable for research, and they are now available for request through the Census Bureau's restricted data research program. This paper discusses the provenance and composition of the files, assesses the coverage and quality of the data, and discussed potential uses of the data for research.

**Keywords:** Taxation, Administrative Data
**JEL Classification Codes:** C81; C55

This paper describes a series of linkable individual-level data from late-twentieth century federal income tax returns. The Internal Revenue Service (IRS) Form 1040 microdata files were held at the Census Bureau since originally being delivered by the IRS shortly after each year's tax returns were processed. These data were not retained at the IRS, and until recently the Census Bureau used them only to create estimates in support of the General Revenue Sharing program and for operational purposes related to the agency's Population Estimates Program. In late 2016, the Census Bureau made de-identified versions of the data available for request by Census Bureau research staff and their collaborators. More recently, the IRS Form 1040 files from 1969 through 1994 were returned to the IRS, and since then the Census Bureau, in collaboration with IRS, has made these data available for use on approved projects through their restricted-use data program.[1]

We discuss the provenance and composition of the files, as well as several important issues for potential users to consider. We then describe data access as these data are highly restricted and currently can only be accessed through secure systems maintained by the Census Bureau and the IRS. Each agency manages a rigorous application process that requires researchers to specify how their work benefits the core goals of these agencies. We conclude by describing some ongoing projects that demonstrate the data's potential. As will become clear, these data are an extraordinarily valuable "new" resource, effectively providing a longitudinal data series covering hundreds of millions of American taxpayers in the late twentieth century.

## Provenance of the historical tax files

The IRS Form 1040 files were derived from the IRS's Individual Master File (IMF) from 1969, 1974, 1979, 1984, 1989, and 1994. The IRS provided the 100-percent IMF extracts on hundreds of

[1] The 1969, 1974, 1979, 1984, 1989, and 1994 IRS Form 1040 files can be requested for use through Census Bureau's restricted data program and the Federal Statistical Research Data Centers (FSRDCs) using the standard application process (SAP) portal. The files are titled, "IRS Form 1040 External Commingled File".

reels of magnetic tape to the Census Bureau for calculations of population and income for the General Revenue Sharing program (National Archives and Records Service, 1988). The State and Local Fiscal Assistance Act of 1972 created the General Revenue Sharing program and defined key terms such as "population" and "income" to be calculated using methods consistent with those "determined by the Bureau of the Census for general statistical purposes" (p. 928). At the start of the program's implementation, there were no official annual population or income estimates below the county level. For this reason, staff in the Census Bureau's Population Division initiated an annual estimation process for the nation's approximately 36,000 cities, towns, townships and other county subdivisions by adjusting estimates from the 1970 Census with the IRS Form 1040 data. The aggregate estimates calculated through these efforts became a resource for researchers, but the individual-level microdata files were retained for use by the Census Bureau only (U.S. Census Bureau, 1984).

Beyond the initial General Revenue Sharing use, the Census Bureau has used these data in additional ways. Between the 1970s and 1990s, staff in the Population Division used these data to produce five-year migration estimates and to create comparisons with decennial censuses (Starsinic, 1974) and to estimate income (Herriot, 1974). Census Bureau researchers routinely use administrative records to supplement census and survey data and to improve the quality of their programs and statistics, and data from IRS has been important to several additional Census Bureau efforts (e.g., Childers & Hogan, 1984; O'Hara, 2006; Sheppard et al., 2013). In 2008, the Population Division delivered the historical IRS Form 1040 files to the Census Bureau's Data Integration Division for permanent storage. The Data Integration Division, which later became the Center for Administrative Records Research and Applications, prepared the files for use by researchers in 2016. The Census Bureau retained "research-ready" versions of the IRS Form 1040 files (described more below); they also returned the version of the processed files to the IRS in 2022 that did not include any variables added by the Census Bureau.

While the data were in Census Bureau custody from their time of delivery until now, the only remaining documentation consists of record layouts and file counts that the Census Bureau's Population Division produced to support transferring the data to the Data Integration Division in 2008. We have not located any original IRS-provided documentation that may have originally accompanied these data. The existing documentation and published reports (e.g., Starsinic, 1974; Herriot, 1974) make clear that the Census Bureau processed the data in various ways to support operational work related to population and income estimates. For instance, we can see that the Census Bureau added various geography variables needed to measure county-to-county migration rates. We have no documentation on these transformations or on any other changes that may have been made to the files over the decades.

Given our limited understanding of how the files were edited by the IRS prior to delivery or by the Census Bureau between 1970 and 2008, we take steps here to document the contents of the files and to provide potential data users with some assurance that the files truly are what they seem to be. These investigations lead us to believe that the historical IRS Form 1040 files indeed contain a limited set of variables from the vast majority of tax returns submitted during each year, as well as variables constructed by the Census Bureau that enhance the data's usability for researchers.

## Contents of the historical IRS Form 1040 files

The historical IRS Form 1040 files consist of a single annual file from each of the years 1969, 1974, 1979, 1984, 1989, and 1994. Each annual file contains between 75 million records (in 1969) and 110 million records (in 1994). Each record represents a single tax return containing a mailing address and personally identifying information (PII) for a tax filer. From 1974 through 1989, there is also PII for the spouse of the tax filer for returns filed jointly, and in 1994 there is PII for filers, spouses, and up to four dependents. In 1969, the filer PII consists of Social Security Number (SSN) only. In 1974 through 1994, the PII includes the SSNs of the filer and the spouse, as well as the filer's name and the spouse's name. In 1994, the data also contain names and SSNs for up to four dependents. In addition to this PII, the datasets contain variables detailing income, the exemptions claimed across each of several categories, and binary variables indicating whether several supplemental schedules were filed. All of the IRS Form 1040 records contain variables indicating how many dependents were included on each return.

### *Variables and coverage*

Table 1 shows the range of variables included in the files and indicates when an item is only available for a subset of years. In general, the items provide much of the detailed information about filers and the income reported on the main 1040 form. The 1969 data have the fewest variables, often with values that are rounded. The 1974 variables have a bit more information, and nearly all variables are present from 1979-1994. There are typically three or more versions of each geography variable available, and the Census Bureau has recently added additional variables to identify detailed geographic areas and to support record linkage (we describe geography variables in more detail below).

**Table 1.** Variable Availability in IRS Form 1040 Files, 1969-1994

Filer information
- Mailing address
- Filing status
- Filer SSN
- Spouse SSN (1974-1994)
- Filer name and spouse name (1974-1994)
- Dependents 1-4 SSNs and names (1994)
- Filer or spouse over 65 (1974-1994)
- Filer or spouse blind (1974-1994)

Income amounts
- Adjusted gross income
- Wage and salary income
- Interest income
- Dividend income
- Gross rents and royalties income (1979-1994)
- Earned income credit (1979-1994)
- Advanced earned income credit (1979-1994)
- Social Security income (1994)
- Tax-exempt interest income (1994)

Geography
- State, county, place, zip code
- Geo-coordinates and other geographies available for some cases (see text)

Exemptions counts
- Spouse
- Children
- Parents
- Age 65 and up
- Children with EIC (1984-1994)
- Deceased (1984-1994)

Schedule filing indicators
- Schedule F filed
- Schedule D filed (1974-1994)
- Schedule C filed (1974-1994)
- Schedule E filed (1974-1994)
- Schedule SE filed (1974-1994)
- Estate return (1989-1994)
- Social Security income (1989-1994)
- Schedule A filed (1994)

It is important to note that these variables do not include every item on IRS Form 1040 and include almost no information from supplemental schedules. For example, while all returns have a variable indicating whether Schedule F (the farm schedule) was filed, none have any information on whether the filer reported on that schedule. The files also lack information on some variables that are present on the first and second pages of Form 1040, such as occupation and taxes paid. We suspect that these items were never delivered to the Census Bureau. Even in the 2020s, when tax return data continues to be an essential component of Census Bureau frames and survey work, the Census Bureau only receives the subset of variables.

Table 2 shows the number of tax returns in each file, as well as the number of individual tax filers, spouses, and dependents who were included in the records. Every tax return contained a primary filer, and for returns that married couples filed jointly, information is also available for the spouse. In 1994, identifying information up to four dependents per tax unit could be included on a return. Thus, for households with more than four dependents as indicated by the exemption counts, they only listed four.

**Table 2.** Number of Records in the IRS Form 1040 Files

| Tax Year | Tax Returns | Filers | Spouses | Dependents |
|---|---|---|---|---|
| 1969 | 75,070,000 | 75,070,000 | - | - |
| 1974 | 80,960,000 | 80,960,000 | 44,950,000 | - |
| 1979 | 90,760,000 | 90,760,000 | 45,610,000 | - |
| 1984 | 94,790,000 | 94,790,000 | 45,910,000 | - |
| 1989 | 108,600,000 | 108,600,000 | 48,710,000 | - |
| 1994 | 110,700,000 | 110,400,000 | 47,300,000 | 69,960,000 |

Note: This includes number of filers, spouses, and dependents with PII listed on the IRS Form 1040 returns. Results were approved for release by the U.S. Census Bureau, approval #CBDRB-FY24-ADEP001-001, CBDRB-FY24-ADEP001-006, CBDRB-FY24-ADEP001-007.

*Additional variables in research-ready files*

In order to prepare the files for use by researchers, the Census Bureau created “research-ready” versions of the data in 2016. These versions have additional variables to facilitate linkage of individuals and addresses between the tax files and to different kinds of data (such as census records), as well as more detailed geographic identifiers than were available on the original data. Since the research-ready files are a Census Bureau product, the additional variables described below were not included as a part of the historical IRS Form 1040 files that were returned to the IRS in 2022.

The research-ready files strip PII and when possible, replace filers’ and spouses’ PII with a “Protected Identification Key” (PIK). PIKs are unique and persistent identifiers that can be used to link an individual across any set of files that have been assigned PIKs. For instance, a researcher working on an approved project could link a respondent’s tax record to their decennial census record, using the census data to add variables such as race or place of birth (items not collected by the IRS). The Census Bureau assigns PIKs by linking records to a composite of existing administrative records. In the case of tax records, this typically involves linking the records to the composite using SSNs and names, and then replacing those variables with a PIK. As can be seen in Table 3, nearly all tax filers were assigned a PIK, as were the vast majority of spouses and dependents in 1994.

**Table 3.** Percentage of Filers, Spouses, and Dependents Assigned Protected Identification Keys (PIKs) in the IRS Form 1040 Files

| Tax year | Filers | Spouses | Dependents |
|---|---|---|---|
| 1969 | 100.0 | - | - |
| 1974 | 100.0 | 96.0 | - |
| 1979 | 99.5 | 96.9 | - |
| 1984 | 99.5 | 86.1 | - |
| 1989 | 98.4 | 84.4 | - |
| 1994 | 99.7 | 97.3 | 94.5 |

Note: Results were approved for release by the U.S. Census Bureau, approval #CBDRB-FY24-ADEP001-001, CBDRB-FY24-ADEP001-006, and CBDRB-FY24-ADEP001-007.

Table 4 shows the number of individuals that were linked from one tax file to another via PIKs, and from any of the tax files to the 1940 and 2000 censuses. The figures on the gray diagonal show the number of individuals that have been assigned unique PIKs in each file. For instance, Table 4's upper left-hand cell shows that 53,360,000 persons were assigned PIKs in the 1940 Census. Cells below the diagonal show the number of individuals with a PIK that are present in the join of any two files. For instance, the second cell down from the top left corner shows that 24,530,000 cases from the 1969 tax data can be linked to the 1940 Census. As this table makes clear, the historical tax files have enormous potential to support research on how individuals' lives have changed over the twentieth century, particularly when combined with other files that have been assigned PIKs.

**Table 4.** Individuals Assigned a Protected Identification Key (PIK) that can be Linked Across Files and to Selected Decennial Censuses

| | 1940 Census | 1969 Tax Data | 1974 Tax Data | 1979 Tax Data | 1984 Tax Data | 1989 Tax Data | 1994 Tax Data | 2000 Census |
|---|---|---|---|---|---|---|---|---|
| 1940 Census | 53,360,000 | | | | | | | |
| 1969 Tax Data | 24,530,000 | 75,070,000 | | | | | | |
| 1974 Tax Data | 37,280,000 | 63,110,000 | 122,700,000 | | | | | |
| 1979 Tax Data | 34,380,000 | 56,680,000 | 106,500,000 | 133,800,000 | | | | |
| 1984 Tax Data | 30,520,000 | 50,150,000 | 94,110,000 | 111,400,000 | 133,800,000 | | | |
| 1989 Tax Data | 26,920,000 | 45,160,000 | 86,460,000 | 103,300,000 | 113,200,000 | 146,200,000 | | |
| 1994 Tax Data | 23,140,000 | 40,870,000 | 80,580,000 | 98,060,000 | 106,000,000 | 126,000,000 | 212,400,000 | |
| 2000 Census | 23,680,000 | 40,070,000 | 80,460,000 | 96,750,000 | 102,300,000 | 118,200,000 | 179,000,000 | 238,900,000 |

Note: the gray diagonal shows the number individuals in each file that received a PIK. Results were approved for release by the U.S. Census Bureau, approval #CBDRB-FY24-ADEP001-001 and CBDRB-FY24-ADEP001-006.

The research-ready files also contain new geographic variables that were not included in the original data or the data files delivered to Data Integration Division in 2008. Using the street addresses from the tax data, the Census Bureau conducted a linkage to the Master Address File (MAF) (Onora & Winkelmann, 2018). The MAF is a constantly updated list of known residential addresses since 2000; it serves as the geographic frame for many census and survey operations (Sullivan & Genadek, 2024). Whenever a tax address was linked to a MAF address, the Census Bureau assigned a "MAFID" variable that identifies the record's county, tract, block, geographic coordinates, and a range of other geographic variables. For many records, even if the linkage process could not assign an exact MAFID, the process was still able to assign the county, tract, and block variables.

The MAF variables have two key advantages over the original geographic variables included in the tax files. First, the MAF variables are available at a more granular level of detail, including precise geolocation. The original variables generally include only contemporaneous Census Place, Census Minor Civil Division (MCD), County, and State. Second, the MAF variables are applied using consistent boundaries over time for counties, tracts, and other areas, using the boundaries established for the 2010 Census. The key limitation of the MAF variables is that they are not available for every record in the historical tax files. Since the historical tax data are decades older than the MAF, many of the tax records that have a complete street address still cannot be associated with a MAF record because that domicile may no longer exist or changes in street naming and ZIP codes make matching the MAF record impossible. In addition, many tax records have only a PO Box or a rural route number and therefore cannot be assigned a precise location.

### Data integrity

Given the limited information that we have about how the files have been managed or modified over the decades, we did several basic integrity checks to assess the extent to which the files are complete and accurate. Below we investigate whether the files contain the expected number of records and how accurate the variable values are. We conclude by highlighting several challenging issues with geography and variable harmonization.

*Data integrity: file coverage.*

We have two ways to assess whether the files have the number of cases that would be expected in a full-universe database of tax returns. First, we verify that the number of records present in each year matches the number of records that the Census Bureau documented for their internal delivery of files from the Population Division to the Data Integration Division in 2008. While limited, this is the only documentation we have about these files. Second, we draw on the multitude of case counts and statistics that the IRS produced after each tax year, comparing the number of records in each year of data to other information that the IRS published about the number and distribution of late-twentieth-century tax records in the corresponding year.

When the historical tax files were transferred internally in 2008, corresponding documentation included the number of records in the 1979, 1984, and 1989 files. Each file was delivered in 132 parts called “cuts”, where each cut included all records within a given range of SSNs. The documentation includes the total number of records as well as the number included in each cut. We have no documentation on the number of records transferred in the 1969 and 1974 files.

In 1984 and 1989, the microdata files contain the exact number of records reported in the internal documentation. In 1979, however, the microdata files have 540,000 *more* records than the internal transfer documentation indicates should be present. The documentation for the 1979 file indicates that one of the cuts delivered by the IRS in 1980 seemed to erroneously *exclude* about 530,000 records that the Census Bureau expected to receive, representing about 0.6% of all tax returns. If these records were in fact delivered to the Census Bureau in 1980 (or later) and were included in the internal transfer in 2008, this would account for most of the additional records that we observe in the currently-held data.

With this exception, the number of records currently available matches the number that the Census Bureau transferred internally in 2008.

Our second method for validating the record counts is to compare the historical tax files to published counts that the IRS's Statistics on Income division (SOI) produced after each year's tax returns were collected. The published counts draw on the full IMF, which is the source of the extracts that IRS provided to the Census Bureau. Table 5 compares the historical tax files' case counts to those from published statistics. The historical tax files at the Census Bureau contain 0.3% to 5% fewer filers than the published counts. The gap is substantially smaller in 1969, 1974, and 1979 (0.3%, 1.8%, 2.1% respectively) than in 1984, 1989, and 1994 (around 5%). Table 6 also shows total Adjusted Gross Income (AGI) published in the SOI reports and the AGI as calculated from the files at the Census Bureau.[2] In every year, the summed AGI of the returns in the historical tax files is between 0.7% and 6.1% less than the total AGI in tables published by the IRS.

**Table 5.** Number of Records in the IRS Form 1040 Files and IRS Publications

| Tax Year | Tax Returns in the IRS Form 1040 Files | Tax Returns in the IRS Publications | IRS Form 1040 File Returns as a percentage of IRS Publications Returns |
|---|---|---|---|
| 1969 | 75,070,000 | 75,280,381 | 99.7% |
| 1974 | 80,960,000 | 82,469,288 | 98.2% |
| 1979 | 90,760,000 | 92,694,302 | 97.9% |
| 1984 | 94,790,000 | 99,752,249 | 95.0% |
| 1989 | 108,600,000 | 113,242,080 | 95.9% |
| 1994 | 110,700,000 | 116,466,422 | 95.0% |

Note: Sources include Statistics of Income, 1972; Statistics of Income, 1977; Statistics of Income, 1982; Statistics of Income, 1986; Statistics of Income, 1991; Statistics of Income, 1996. Results were approved for release by the U.S. Census Bureau, approval #CBDRB-FY24-ADEP001-001 and CBDRB-FY24-ADEP001-006.

[2] The IRS published statistics from 1979 were based on a sample of tax return data rather than analyses of the full files produced in 1969, 1974, 1984, 1989, and 1994.

**Table 6.** Total Adjusted Gross Income (AGI) in the IRS Form 1040 Files and IRS Publications

| Tax Year | Sum of AGI Values in the IRS Form 1040 Files (in 1000s) | Sum of AGI Values in the IRS Publications (in 1000s) | IRS Form 1040 AGI as a Percentage of IRS Publications AGI |
|---|---|---|---|
| 1969 | $597,800,000 | $602,301,527 | 99.3% |
| 1974 | $882,100,000 | $896,182,659 | 98.4% |
| 1979 | $1,444,000,000 | $1,465,394,530 | 98.5% |
| 1984 | $2,043,000,000 | $2,134,035,012 | 95.7% |
| 1989 | $3,129,000,000 | $3,250,669,292 | 96.3% |
| 1994 | $3,660,000,000 | $3,898,339,504 | 93.9% |

Note: Sources include Statistics of Income, 1972; Statistics of Income, 1977; Statistics of Income, 1982; Statistics of Income, 1986; Statistics of Income, 1991; Statistics of Income, 1996. Results were approved for release by the U.S. Census Bureau, approval #CBDRB-FY24-ADEP001-001, CBDRB-FY24-ADEP001-006, CBDRB-FY24-ADEP001-007.

We have not been able to determine why the historical tax files have fewer tax returns and less total AGI than are reported in IRS publications. We suspect that many of the "missing" returns are from late filers. We used more detailed published reports to investigate whether these gaps were due to the omission of people in particular states or regions. In the years with the largest differences, from 1979-1989, the vast majority of states in the published data reported more returns than are in the microdata. The largest discrepancies were in Alaska (published tables had about 10% more cases than the historical tax files in 1979-1989), California (8-10% more in 1984-1989), and the District of Columbia (8% more in 1984-1989). Other states had peaks of 5%-7% in given years. Even with these outliers, there is fluctuation in both directions—some states have *fewer* records in the published statistics. The average absolute value of these state level differences ranges from 1.1% in 1969 to 4.4% in 1984. We suspect that one key reason for the state-to-state variation relates to how non-standard returns are included in published statistics. Military, international, and erroneous tax returns and even returns from the District of Columbia are incorporated into various states' totals, and the level of documentation on the methods used in these processes is not always clear or complete.

*Data integrity: variable accuracy*

In addition to doing basic comparisons on the number of records in the IRS 1040 microdata at the Census Bureau, we also attempt to assess the accuracy of the values available on each case. We did this by conducting an individual-level linkage of filers and spouses from the historical tax files to their records from nearby years in the Current Population Survey (CPS). Conducted by the Bureau of Labor Statistics and the Census Bureau, the CPS is a monthly survey with microdata available since the early 1960s. We used CPS cases associated with the Annual Social and Economic Supplement (ASEC), which during this period generally contained about 130,000 to 150,000 cases. We used the CPS ASEC because these are among the very few historical CPS samples that have been assigned PIKs, and even then, only for a very limited set of years. For this reason, we use the closest possible CPS ASEC file that we can for each tax year. Thus, our analysis focuses on the CPS ASEC sample from 1973 (matched to the 1969 and 1974 tax data), 1985 (matched to the 1984 tax data), and 1991 (matched to the 1989 tax data). Table 7 shows the number of cases assigned a PIK in each year's CPS sample, as well as the number and percentage of those cases that we were able to assign a PIK to in the relevant tax file.[3]

[3] There are a number of potential explanations for why these linkage rates are not higher. First, recall that the tax data from 1969-1989 exclude dependents; each return has a primary filer, and some also have a spouse. Well over a quarter of the population was aged 18 and under during this period. Since children are present in the CPS data and not in the tax data (unless they were on a tax return as the primary filer or spouse), almost all children with a PIK in the CPS are not linkable to the tax data. In 1969, when the linkage rate between tax data and CPS data is particularly low, spouses were also not present in the tax data and thus are not linkable to the CPS. Finally, PIK assignment is a probabilistic process that always includes some error, so we would not expect perfection in any case. These linkage rates seem reasonable for a linkage of a full-universe tax file to contemporaneous survey data.

**Table 7.** Number of Records Linked Between the Current Population Survey (CPS) and the IRS Form 1040 Files

| | CPS Records with PIKs | CPS Records Linked to IRS Form 1040 Files | CPS Records Linked to IRS Form 1040 Files |
|---|---|---|---|
| 1973 CPS to 1969 tax | 84,000 | 38,000 | 45.2% |
| 1973 CPS to 1974 tax | 84,000 | 67,000 | 79.8% |
| 1985 CPS to 1984 tax | 64,500 | 40,000 | 62.0% |
| 1991 CPS to 1989 tax | 106,000 | 78,500 | 74.1% |

Note: Results were approved for release by the U.S. Census Bureau, approval #CBDRB-FY24-ADEP001-001.

We performed three analyses to assess the accuracy of the values in the tax data based on the information in the CPS for the respondent with the same PIK. First, we first compared the marital status in the tax data for couples that were also in the CPS. We started by restricting our analysis to CPS respondents who were identified as either a head of household or that person's spouse. While other household members have an indicator of whether or not they are married, it is often impossible to determine whether they were married to another member of the household, so we excluded unmarried household heads and everyone else in the household other than the head of household and named spouse named spouse.

Focusing on CPS household heads and their spouses who both had PIKs in 1985, we linked these individuals to the 1984 tax data using the PIKs. Whenever we linked a CPS head to a filer or spouse who was part of a couple in the tax data, we noted when we were observing the same married couple in the two files. This process allowed for the CPS head and spouse to be in either position in the tax data. For instance, the CPS "spouse" could be the primary filer in the tax data and the CPS "head" could be the spouse in the tax data. There were 22,000 individuals who were part of a head/spouse pair with PIK values in the 1985 CPS sample. When we linked these people to filer/spouse pairs with a PIK in the 1984 tax data, the PIK for the CPS couple matched the PIK for the tax couple 99% of the time. This finding gives us a high degree of confidence that, when PIKs are available, the historical IRS

Form 1040 files accurately identify spousal pairs with a similar degree of accuracy as is observed in high-quality survey data from the same era.

Our second effort to validate the accuracy of variables in the tax data involved linking every possible individual between the CPS and IRS Form 1040 data (not just household heads and spouses), and then comparing marital status values on the files. This approach allows us to consider a broader range of people than the above, since we can include all of those who are single, as well as married people who were not the household head or spouse. Recall that we had to exclude the latter group from our first analysis because we could not determine to whom they were married or whether their spouse was in the household.

We conducted this analysis by first creating a binary married/unmarried variable for all cases having a PIK in the tax data and the CPS data. In the tax data, we observe marital status in the filing status variable. We coded the following statuses as married: “married filing jointly,” “married - filing separately,” and “separate claiming spouse.” We coded the following statuses as single: “single filing,” “unmarried head of household,” and “surviving.” In the 1989 CPS data, to take one example, we coded “married civilian spouse present,” “married - armed forces spouse present,” “married - spouse absent,” and “separated” as married. We coded “widowed,” “divorced,” and “never married” as unmarried. While other years of CPS data had slightly different values for the marital status variables, it was always clear enough whether to code each as married or unmarried in similar ways to 1989.

Table 8 shows the correspondence between our indicators of marital status in the tax data and the CPS. For all years, between 88% and 97% of respondents had the same marital status value in the two files. Since there were one to three years between the CPS and tax data, it is not surprising that 3% to 12% of cases had different marital status values in the two files. The one exception is the 1984 tax data and the 1985 CPS data, which were both collected around the Spring of 1985 (though tax filers should have reported their marital status as of the previous year). This particular comparison had the most correspondence between the two files, with 97% of cases having the same marital status values. The remaining differences are likely explained by changes in marital status, as well as linkage error. The relatively high number of married persons in the 1969 tax data to 1973 CPS match derives from a decision to include only persons who were household heads in the CPS.

In our third and final effort at variable validation, we conducted the same type of analysis using the age variable. While an actual age value is not present in the IRS Form 1040 data, we do have an indicator of whether a person is aged 0-64 or 65+. For 1979-1994, we have separate variables indicating whether the filer is 65+ and whether the spouse is 65+. For the 1969 and 1974 tax data, this measure is even more crude. For those years, the file contains a single binary variable indicating when either the filer or spouse is either 65+ or blind. Since the flag in these years only indicates that either the filer or spouse (or both) are 65+ or blind, we do not expect anything near a perfect match. We created similar age indicators for the CPS data, adding or subtracting years as necessary. For example, we add one year to reported age in the 1973 CPS when comparing to the 1974 tax data. Table 9 shows the results of this age comparison. Only 1%-4% of cases differ in this crude measure of age.

**Table 8.** Marital Status of Linked Records in the Current Population Survey (CPS) and the IRS Form 1040 Files

| | 1969 Form 1040 and 1973 CPS File | 1974 Form 1040 and 1973 CPS File | 1984 Form 1040 and 1985 CPS File | 1989 Form 1040 and 1991 CPS File |
|---|---|---|---|---|
| Married in both files | 84% | 57% | 64% | 63% |
| Unmarried in both files | 6% | 32% | 32% | 26% |
| Marital Status Differed | 10% | 11% | 3% | 12% |
| Total | 100% | 100% | 100% | 100% |
| Number of cases | 26,530 | 46,550 | 40,500 | 68,000 |

Note: Results were approved for release by the U.S. Census Bureau, approval #CBDRB-FY24-ADEP001-001 and CBDRB-FY24-ADEP001-006.

**Table 9.** Age of Linked Records in the Current Population Survey (CPS) and the IRS Form 1040 Files

| | 1969 Form 1040 and 1973 CPS File | 1974 Form 1040 and 1973 CPS File | 1984 Form 1040 and 1985 CPS File | 1989 Form 1040 and 1991 CPS File |
|---|---|---|---|---|
| Age 65 and Above in both files | 6% | 8% | 10% | 11% |
| Under Age 65 in both files | 92% | 90% | 89% | 85% |
| Age Range Differed | 2% | 2% | 1% | 4% |
| Total | 100% | 100% | 100% | 100% |
| Number of cases | 26,530 | 46,550 | 40,500 | 68,000 |

Note: Results were approved for release by the U.S. Census Bureau, approval #CBDRB-FY24-ADEP001-001 and CBDRB-FY24-ADEP001-006.

## Further issues for user consideration

There are several additional issues that researchers should be aware of when using the historical IRS Form 1040 files. As mentioned above, there are several different versions of geography variables associated with the historical tax files. These include the original address fields from the tax returns, additional geographic variables created by the Census Bureau to support operational needs related to population estimates and revenue sharing, and yet more geographic variables created by the Census Bureau to support research (as part of the research-ready files). The original mailing address includes a city, state, and ZIP code. The Census Bureau's population estimates group derived the additional variables using various intricate statistical methods, although the technical details of how these variables were constructed have been lost (Peil, personal communication, September 7, 2022). Finally, the research-ready files have *additional* geography variables generated from the Master Address File (MAF) linkage (including geocordinates).

For these reasons, the historical IRS Form 1040 files have varying numbers of geographic variables (ranges shown in parentheses) identifying state (4-8), county (2-7), ZIP code (1-3), city/place (3-8), and minor civil division (2-5). These variables have differing degrees of coverage across records, and in some cases the different variables contain different values (even for major geographic designations such as county or state). The large number of geographic variables creates many opportunities (as well as potential confusion) for researchers interested in spatial analysis. We analyzed, compared, and investigated these variables in order to determine what they are, from where they originate, which ones to use for different purposes, and how they compare to summary data published by the IRS.

Due to the sheer number of geography variables, we developed recommendations for future researchers, depending upon the needs of the project. For city, state, ZIP code, and/or street address, we recommend using the original mailing address. For point-level geocoding, or to identify consistent 21st century counties, tracts, or blocks, we recommend using the MAF-based variables that the Census

Bureau created for the research-ready files. For county codes assigned contemporaneously to the Form 1040 tax year (e.g., the 1970 census county boundaries), researchers should use the “probability” county variable. For contemporaneous census tract data, see the resources we developed to assign historical census tracts (Bleckley, Genadek, & Alexander, 2023a).

While the selection of geography variables is one of the more complex issues that users are likely to encounter with these files, there are additional challenges that relate more to these being large historical files. The data values are all presented as strings; thus, any analysis of numeric variables requires reformatting appropriate to the statistical software being used. The files were created during a time when digital storage was expensive, so every byte mattered. To save space, in 1969 and 1974, income data are split into characteristic and mantissa variables, where the mantissa can be multiplied by 10 raised to the characteristic variable value to calculate the income variable. Users of these data should note that in the 1969 data file, the variable labels for the mantissa and characteristic have been transposed; thus, the variable labeled "characteristic" should be treated as the mantissa and vice versa. Additionally, in 1969 and 1974 the Adjusted Gross Income (AGI) has a sign variable to designate whether the AGI is positive or negative; however, in 1969 all values for the AGI sign variable are blank, so users cannot determine if any the AGI values are negative. In 1969-1994, the negative sign is embedded in the AGI variable itself.

Many researchers will want to use these files as a time-series or as a longitudinal panel. Because Census Bureau staff used these data at various points across a three-decade period, the variable names and labels are usually inconsistent from year to year. Researchers on projects approved to use these data can request CES Technical Note Series #23-11 (Bleckley, Genadek, & Alexander, 2023b) to access harmonization programs to make variable names consistent over time.

## Recent projects using the historical IRS Form 1040 data

As stated above, these data have been available for analysis by Census Bureau researchers and their external colleagues since 2016. Non-governmental researchers can now propose projects using these data through the Census Bureau restricted-use data program.[4] The historical IRS Form 1040 files are a useful panel on their own, and their value can be increased even further by preparing them for longitudinal use and linking them to other administrative records, or surveys from prior or future decades. Several researchers have already incorporated these tax data into their work, and we present just a few of these to showcase the wide range of possibilities these data bring to research.

The tax data allow researchers to look at economic phenomena in the U.S. with great accuracy at a wide variety of geographies as well as at the individual level over time. Rinz and Voorheis (2023) link the tax data to Social Security Administration (SSA) records to examine income distribution over time. They find that the highest end of the income distribution has become increasingly dissimilar across the past four decades, a discovery that provides new nuance to the literature around income convergence. Other researchers use the tax data to look even further back in history. Massey and Rothbaum (2021) measure intergenerational economic opportunity using 1940 Census data linked to the 1974 and 1979 tax records. They find key differences in children's outcomes based on race, geography, and family structure. Linking those same datasets to 2000 Census data allows for granular spatial analyses, exploring the negative long-term impacts of redlining in the 1930s (Aaaronson, Hartley, Mazumder, & Stinson, 2022).

The tax data are also being incorporated into Census Bureau infrastructure projects that intend to create anonymized linked longitudinal data for use by future researchers. The National Experimental Wellbeing Statistics project brings together census, survey, commercial, and administrative data (including the historical tax data) in order to overcome challenges in measuring income and poverty (Bee et al., 2023). This creates a powerful new tool to explore economic wellbeing and inequality with

[4] More information on applying to use these data is available here: https://www.census.gov/topics/research/guidance/restricted-use-microdata/standard-application-process.html

increased accuracy. The Decennial Census Digitization and Linkage project is in the process of recovering name data for the 1960-1990 censuses to assign PIKs to these records, allowing individual-level longitudinal analyses of these de-identified censuses for the first time (Genadek and Alexander, 2019; Alexander and Genadek, 2023). The tax data, alongside other administrative records, are being used to facilitate the assignment of PIKs. Projects like these allow researchers to focus on developing novel research questions and robust analyses, rather than data cleaning and linkage.

## Conclusion

IRS Form 1040 microdata from 1969, 1974, 1979, 1984, 1989, and 1994 offer researchers hundreds of millions of longitudinally-linkable, de-identified individual records, providing new potential for innovative discoveries on the social and economic issues of late-20th century American life. Efforts by the Census Bureau to add value through data linkage and geocoding have developed research-ready datasets, facilitating longitudinal research at the individual level, with granular geographies, and with the potential to link these files to other administrative records and survey data. We have determined that these data are nearly complete extracts of all tax returns filed in these years. While the number of returns in a given year is up to 5% less than published counts, this difference does not seem to originate from a particular geography, income level, or age group. These data are the most complete and accurate set of individual tax records available to researchers, and efforts by the Census Bureau and IRS have created a new opportunity to access these rich and important datasets.

## References

Aaronson, D., Hartley, D., Mazumder, B., & Stinson, M. (2022). The Long-run Effects of the 1930s Redlining Maps on Children. CES Working Paper 22-56. https://www2.census.gov/ces/wp/2022/CES-WP-22-56.pdf

Alexander, J. T. & Genadek, K. R. (2023). Using administrative records to support the linkage of census data: protocol for building a longitudinal infrastructure of U.S. census records. *International Journal of Population Data Science*, *7*(4). https://doi.org/10.23889/ijpds.v7i4.1764

Bee, A., Mitchell, J., Mittag, N., Rothbaum, J., Sanders, C., Schmidt, L., & Unrath, M. (2023). National Experimental Wellbeing Statistics: Version 1. CES Working Paper 23-04. https://www2.census.gov/library/working-papers/2023/adrm/ces/CES-WP-23-04.pdf

Bleckley, D. A., Genadek, K. R., & Alexander, J. T. (2023a). Assigning Contemporaneous Census Tracts to Historical Income Tax Data. ADEP Working Paper 2023-03. https://www.census.gov/content/dam/Census/library/working-papers/2023/econ/censustract_taxdata.pdf

Bleckley, D. A., Genadek, K. R., & Alexander, J. T. (2023b). Harmonization of Variable Names in 1969-2000 Internal Revenue Service Individual Form 1040 Data. CES Technical Notes Series 23-11, Center for Economic Studies, U.S. Census Bureau. Harmonization of Variable Names in 1969-2000 Internal Revenue Service Individual Form 1040 Data (repec.org)

Childers, D. R. & Hogan, H. R. (1984). The IRS/Census Direct Match Study. SRD Research Report Number: CENSUS/SRD/RR-84/11. https://www.census.gov/content/dam/Census/library/working-papers/1984/adrm/rr84-11.pdf

Finlay, K., & Genadek, K. R. (2021). Measuring all-cause mortality with the Census Numident file. American journal of public health, 111(S2), S141-S148. https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2021.306217

Genadek, K. R., & Alexander, J. T. (2019). The Decennial Census Digitization and Linkage Project. ADEP Working Paper 2019-01. https://www.census.gov/content/dam/Census/library/working-papers/2019/econ/dcdl-workingpaper.pdf

Herriot, R. A. (1974). Preparation of Final Revenue Sharing Estimates of Money Income for Political Jurisdictions. Statistical Methodology of Revenue Sharing and Related Estimates Studies. Census Tract Papers Series GE-40, No. 10 pp. 18-25. Washington, DC: U.S. Government Printing Office. https://hdl.handle.net/2027/mdp.39015086981753?urlappend=%3Bseq=574%3Bownerid=113767829-580

Massey, C. & Rothbaum J. (2021). The Geography of Opportunity over Time. SEHSD Working Paper 2021-23. https://www.census.gov/content/dam/Census/library/working-papers/2021/demo/sehsd-wp2021-23.pdf

National Archives and Records Service. (1988). Request for Records Disposition Authority. Job No. N1-29-87-1. https://www.archives.gov/files/records-mgmt/rcs/schedules/departments/department-of-commerce/rg-0029/n1-029-87-001_sf115.pdf

O'Hara, A. (2006). Tax Variable Imputation in the Current Population Survey. IRS Research Conference. https://www.irs.gov/pub/irs-soi/06ohara.pdf

Onorato, D. & Winkelmann, J. (2018, August 24). Assigning tracts to 1040 forms [Memorandum]. U.S. Census Bureau.

Rinz, K. & Vooheis, J. (2023). Re-examining Regional Income Convergence: A Distributional Approach. CES Working Paper 23-05. https://www2.census.gov/library/working-papers/2023/adrm/ces/CES-WP-23-05.pdf

Sheppard, D., Stewart, T., Rothhaas, C., Lestina, F., Compton E., Machowski, J., & Smith, D. (2013). 2010 Census Administrative Records Use for Coverage Problems Evaluation Report. 2010 Census Planning Memoranda Series, No. 254. https://www.census.gov/content/dam/Census/library/publications/2013/dec/2010_cpex_254.pdf

Starsinic, D. E. (1974). Development of Population Estimates of Revenue Sharing Areas. Statistical Methodology of Revenue Sharing and Related Estimates Studies. Census Tract Papers Series GE-40, No. 10 pp. 2-7. Washington, DC: U.S. Government Printing Office. https://hdl.handle.net/2027/mdp.39015086981753?urlappend=%3Bseq=560%3Bownerid=113767829-566

State and Local Fiscal Assistance Act of 1972, Pub. L. No. 92-512, 86 Stat. 919-947 (1972). https://www.govinfo.gov/content/pkg/STATUTE-86/pdf/STATUTE-86-Pg919.pdf

Statistics of Income. (1972). *Zip Code Area Data: Supplemental 1969*. U.S. Department of Treasury. Internal Revenue Service. https://books.google.com/books?id=yDxXwQEACAAJ

Statistics of Income. (1977). *Small Area Data: Supplemental 1974*. U.S. Department of Treasury. Internal Revenue Service. https://www.irs.gov/pub/irs-soi/74insmallar.pdf

Statistics of Income. (1982). *Individual Income Tax Returns 1979*. U.S. Department of Treasury. Internal Revenue Service. https://hdl.handle.net/2027/uiug.30112002867205

Statistics of Income. (1986). *SOI Bulletin Fall 1986 6*(2). U.S. Department of Treasury. Internal Revenue Service. https://www.irs.gov/pub/irs-soi/86rpfallbul.pdf

Statistics of Income. (1991). *SOI Bulletin Winter 1990-1991 10*(3). U.S. Department of Treasury. Internal Revenue Service. https://www.irs.gov/pub/irs-soi/91rpwinbul.pdf

Statistics of Income. (1996). *SOI Bulletin Spring 1996 15*(4). U.S. Department of Treasury. Internal Revenue Service. https://www.irs.gov/pub/irs-soi/96rpsprbul.pdf

Sullivan J.W., Genadek, K. R., (2024). Using the Census Bureau's Master Address File for Migration Research. ADEP Working Paper 2024-01. https://www.census.gov/content/dam/Census/library/working-papers/2024/econ/mafarf_migration.pdf

U. S. Bureau of the Census. (1994). Geographic Area Reference Manual. U.S. Department of Commerce. https://www2.census.gov/geo/pdfs/reference/GARM/Ch8GARM.pdf

U.S. Census Bureau. (1984). Population and Income Estimates for the United States, 1969-1973. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR00078.v1