ADEP WORKING PAPER SERIES

Assigning Contemporaneous Census Tracts to Historical Income Tax Data

David A. Bleckley
University of Michigan

Katie R. Genadek
U.S. Census Bureau

J. Trent Alexander
University of Michigan

Working Paper 2023-03
August 2023

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

Disclaimer: Any conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. All results were approved for release by the Disclosure Review Board of the U.S. Census Bureau (Data Management System number: P-7506192, Disclosure Review Board (DRB) approval numbers: CBDRB-FY22-ERD002-016 and CBDRB-FY22-ERD002-019).

Assigning Contemporaneous Census Tracts to Historical Income Tax Data

David A. Bleckley, University of Michigan, dbleckle@umich.edu
Katie R. Genadek, U.S. Census Bureau, katie.r.genadek@census.gov
J. Trent Alexander, University of Michigan, jtalex@umich.edu

ADEP Working Paper 2023-03
August 2023

Abstract

This paper presents methods to assign contemporaneous geographic variables to recently recovered historical tax data. We explore various approaches to assign census tract variables to address-level data from the Internal Revenue Service (IRS) from the 1960s-1980s, with the goal of documenting this extraordinary data resource and preparing the files for spatial analyses and linkage to other contemporaneous data. Our efforts to associate mailing addresses with their contemporaneous census tracts were complicated by the changing nature of census tracts, technical limitations on how restricted data can be accessed and transformed, and the lack of year-specific street maps overlayed with year-specific census tract boundaries. We present three methods for assigning contemporaneous census tracts to historical records, evaluate the results, and conclude with recommendations and limitations.

Keywords: Data linkage; income tax data; census tracts; historical data
JEL Classification Codes: C81, H2

Introduction

This paper presents methods of assigning contemporaneous geographic variables to recently recovered historical tax data. We explore various approaches to assign census tract variables to address-level data from the Internal Revenue Service (IRS) from the 1960s-1980s, with the goal of documenting this extraordinary data resource and preparing the files for spatial analyses and linkage to other historical data. Our efforts to associate mailing addresses with historical census tracts were complicated by the changing nature of census tracts, technical limitations on how restricted data can be accessed and transformed, and the lack of year-specific street maps overlayed with contemporaneous census tract boundaries. We present three methods for assigning contemporaneous census tracts to historical records, evaluate the results, and conclude with recommendations and limitations.

Background

Well-defined and -documented geographic units are of vital importance to data analysis and record linkage (Jaro 1978; Perlmann 1979). In order to study populations spatially and temporally, researchers must define spaces, observe populations within those spaces, and understand how both the populations and spatial definitions change and interact. Minimizing or eliminating variability in these definitions facilitates spatial analysis. Thus, efforts are often made to keep geographic boundaries constant over time, with the general goal of census tracts representing "relatively permanent small-area geographic divisions of a county." (U.S. Bureau of the Census, 2018).

The extent of coverage of census tracts changed dramatically during the late-twentieth century. Figure 1 shows how relatively little of the United States' land area was divided into census tracts in 1970, but by 1990 the entire country was assigned a census tract or equivalent "block numbering areas" (U.S. Bureau of the Census, 1995). Between the 1970 Census and the 1990 Census, the number of census tracts increased by 46%, from 34,706 to 50,690 (U.S. Bureau of the Census, 1976; U.S. Bureau of the Census, 1995). Because census tracts were assigned only to more densely-populated areas in earlier years, the increase in the proportion of the population living in census tract-assigned areas did not increase to the same extent. In the 1970 Census, about 72% of the U.S. population lived inside a census tract (U.S. Census Bureau, 1976), and in the 1990 Census all people residing within the U.S. lived in a census tract-assigned area.

Fig. 1 Maps depicting the coverage of census tracts in 1970 (top), 1980 (center), and 1990 (bottom) Sources: (Manson et al., 2021) Software: (Esri, Inc., 2019)

Not only has the overall coverage of census tract-designated land changed, individual census tracts also have evolved. The land area bounded by a census tract designation, the census tract identification number, and the size and demographics of the population living in a tract can all change. To briefly illustrate this

change, we present the census tracts where the Census Bureau headquarters and the Inter-university Consortium for Political and Social Research (ICPSR) offices are located. The census tract identification numbers, boundaries, populations, and demographics of each location's assigned census tract changed from 1970 to 2020 (Table 1). However, the extent of these changes differs between the Suitland, Maryland and Ann Arbor, Michigan examples. The identification numbers of both census tracts changed—in fact, neither of the 1970 census tract numbers still existed in 2020. The Suitland, Maryland census tract was subdivided for the 2000 Census (U.S. Census Bureau, 2000a; U.S. Census Bureau, 2000b) (see Figure 2a), while the Ann Arbor, Michigan census tract boundary changed very little from 1970 to 2020 (Figure 2b). The population increased in the Ann Arbor census tract, while the Suitland census tract decreased dramatically due to the subdivision—it can be inferred that the census tract was subdivided because the population was nearing the 8,000-person maximum (U.S. Bureau of the Census, 2018). As one indicator of demographic change over time, the ratio of Black people to White people increased 100 times in the Suitland census tract from 1970 to 2020, while that ratio remained essentially unchanged in the Ann Arbor census tract.

Table 1 Comparisons of 1970 and 2020 census tracts for Census Bureau Headquarters and ICPSR offices

	Census Bureau Headquarters	ICPSR offices
City, State	Suitland, Maryland	Ann Arbor, Michigan
1970 census tract number [a]	8024.01	0005.00
2020 census tract number [b]	8024.05	4005.00
1970 census tract area (square miles)	1.8	0.38
2020 census tract area (square miles)	1.5	0.39
1970 census tract population [c]	7441	6018
2020 census tract population [d]	3916	7235
1970 ratio of Black:White population [c, e]	0.29	0.05
2020 ratio of Black:White population [d, f]	29.78	0.05

Sources: [a] (Manson et al., 2021); [b] (U.S. Census Bureau, 2022); [c] (U.S. Bureau of the Census, 1972a; U.S. Bureau of the Census, 1972b); [d] (U.S. Census Bureau, 2020)

[e] Calculated by dividing the category "Negro" by "White"

[f] Calculated by dividing the category "Black or African American alone" by "White alone"

Fig. 2 Maps depicting 1970 (solid black) and 2020 (dashed gray) census tract boundaries for locations of Census Bureau Headquarters (a) and ICPSR office (b). Vertical hatched fill indicates 1970 census tract extent. Horizontal hatched fill indicates 2020 census tract extent. Sources: (Manson et al., 2021; U.S. Census Bureau, 2022) Software: (Esri, Inc., 2019)

Census tract boundaries, names, and characteristics can change in all manner of ways over time, making the use of year-specific census tract identification crucial to historical data use and linkage (Lee et al., 2008, Perlmann, 1979). Discordant geographies across time can lead to linkage, matching, and analytic errors (e.g., Raymer et al., 2020). Resources created to facilitate the translation of census tract boundaries over longitudinal datasets allow researchers to utilize census tract-level data despite these changes, minimizing such discordance. The IPUMS National Historical Geographic Information System—NHGIS (Manson et al., 2021) disseminates shapefiles of historical boundaries, including census tracts. Logan, Xu, Stults, and Zhang (2020) have created a longitudinal database of census tract-level social and economic data as well as probabilistic crosswalks, allowing researchers to translate and compare census tract-level data over time. Both of these resources were integral to the implementation of our work. However, neither provides underlying historical street maps for geocoding street addresses.

Our main goal in assigning contemporaneous geographic variables to historical tax data is motivated by the needs of the Decennial Census Digitization and Linkage (DCDL) project. The DCDL project is an initiative to link individual respondent records from the decennial censuses of 1960 through 1990 (Genadek & Alexander, 2019, Genadek & Alexander, 2022). The DCDL project uses administrative records to facilitate the linkage of census respondents' records over time (Alexander & Genadek, 2023). Specifically, when we can make an address-based match between a census record and another federal agency record from the same year, that linkage can serve to validate the census record and enrich it with additional information such as a Social Security Number (which is not collected by the decennial census). Since the 1960 through 1980 census microdata files do not have street addresses but do have county and census tract designations, we need the administrative records to also have reliable county and census tract variables to help make a match that can provide new information about the census respondent.

We describe our efforts both as methodological documentation on the DCDL project and as essential user documentation for others who access these data. The tax files we describe are in use by many research projects at the Census Bureau and have also recently become available through the IRS's Joint Statistical Research Program (https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program). Potential users of those files can use our documentation to plan their research with the low-level geographic variables. In addition to being able conduct analyses using small areas, the contemporaneous tract variables in these data have significant potential to support contextual analyses using the widely varied, tract-level data that the Census Bureau released following each decennial census (available via NHGIS at https://www.nhgis.org/). Our work thus serves as documentation for critical geography variables in this extraordinary data resource.

Data

We assign small geographic area identifiers to IRS Form 1040 data files from 1969, 1974, 1979, 1984 and 1989, which were edited and stored by the U.S. Census Bureau; see Table 2 for record counts by year. The IRS provided these data to the Census Bureau to estimate migration and to determine population counts for federal revenue sharing (Fay & Herriot, 1979, Spencer, 1980). Between the 1970s and 1990s, researchers in the Census Bureau's Population Division (POP) used these data to produce migration estimates and to create comparisons with decennial censuses. In 2008, POP delivered these files to the Census Bureau Data Integration Division for permanent storage (Lamas & Johnson, 2008).

Table 2 Number of records (tax returns) in 1040 data files (rounded), by year

	1969	1974	1979	1984	1989
Record count	75,070,000	80,960,000	90,760,000	94,790,000	108,600,000

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

These data are from IRS Form 1040 individual tax returns, with each record having a mailing address. While most mailing addresses are essentially geographic points, there are significant challenges for using these street addresses for spatial analyses. First, the addresses are free text provided by the person filing their taxes. As a result, they are unstructured and unstandardized. While many records list a standard street address comprised of house number, street name, street type, and direction, there are also many records listing a place or institution name (crossroads, campground, workplace, building name, etc.). Additionally, many records are post office boxes or rural routes. Second, because the records originate as filer-filled forms which were digitized through manual data entry, errors in the address field could have been introduced through initial misspellings, illegibility, and data entry issues. A third challenge relates to geographic changes over time, as street maps and numbering systems have changed both from 1969-1989 and from 1969 to the present.

These issues create additional challenges for point-based spatial analyses. Small geographic unit-based analyses are certainly a viable alternative for many types of research. In fact, the Census Bureau and the IRS both created numerous geographic variables to assist in their own use of these files, though very limited documentation exists about these variables or their creation. This paper provides an overview of the variety of geographic variables available in the Census Bureau-held Form 1040 datasets and describes attempts to create usable census tract assignments.

Available Geographic Unit Variables

The Census Bureau maintains two data files for each year of the IRS Form 1040 records. The first of these files contains all substantive variables, several types of geographic variables, and a Protected Identification Key (PIK). PIKs are anonymized unique identifiers that the Census Bureau creates to facilitate record linkage across files that have been assigned PIKs. The PIKs on these files were assigned through the "Quick PIK" process, which deterministically assigns a PIK based on each record's Social Security Number (Wagner & Layne, 2014). The second file includes geographic variables that the Census Bureau created by matching the IRS street addresses to a 2010 version of the Master Address File (MAF). The MAF is a database of all addresses known to the Census Bureau, along with detailed geographic variables for each address (e.g., block, census tract, metropolitan area, etc.). The "MAF Match" process attempted to associate the pre-1990 IRS addresses with 2010-era street addresses in the MAF (Onora & Winkelmann, 2018, Wagner & Layne, 2014).

We refer to the two sets of files as the "main files" and "geographic supplement files." The main files contain the original IRS variables and additional variables created by the Census Bureau's Population Division, including numerous geographic variables. The geographic supplement files contain only the MAF identification number and other geographic variables created in the MAF Match process, along with a unique identifier that allows users to merge data back to the main files. While the geographic variables in the main files generally refer to current-year geographies (e.g., 1979 county definitions assigned to 1979 records), the variables in the geographic supplement files always refer to 2010-2017 geographies (e.g., 2010 county definitions assigned to 1979 records).

Depending on the year, the main data files have multiple versions of each of the geographic variables, with varying degrees of coverage across records. Many of these variables were created by the Census Bureau, with slight differences to facilitate various methods of population estimation. The geographic variables in the main data files include 4-8 state variables, 2-7 county variables, 1-3 ZIP code variables, 3-8 city/place variables, and 2-5 minor civil division variables. The geographic supplement files have three sets of variables for state, county, census tract, block, and block suffix, as well as one set of latitude and longitude variables. The geographic supplement files increase in coverage over time. For example, 71% and 62% of 1969 records were assigned 2010 census tracts and latitudes/longitudes, compared to 80% and 73% of 1989 records, respectively (see Table 3). Overall, there is a large number of geographic variables in these data files, but with the exception of the census tract, block, and latitude/longitude variables in the

geographic supplement files, none of these geographic variables is granular enough for local-level spatial analyses.

Table 3 Percent of records assigned a 2010 census tract and geographic coordinates in the MAF Match process, by year

1040 Year	Percent of Records Assigned 2010 Census Tracts during the MAF Match process	Percent of Records Assigned a Latitude and Longitude during the MAF Match process
1969	71%	62%
1974	72%	64%
1979	75%	68%
1984	76%	69%
1989	80%	73%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

Since the MAF Match process assigned 2010 census tracts to the pre-1990 tax data, many of these census tracts did not exist at the time when the tax data were collected (see Table 4). The reason for this is a combination of the country being partially divided into census tracts in 1970 and the changes in census tract designations over time. Of the 1969 tax records assigned to a 2010 census tract number by the MAF Match process, only 51% were assigned to census tracts that existed in 1970 (36% of all 1969 records). In comparison, 63% of records assigned to 2010 census tracts were assigned to census tracts that existed in 1990 (51% of all 1989 records). Even when a census tract number did exist previously, it did not necessarily refer to the same geographic area, since the shape of census tracts regularly changed over time. The assignment of historical 1040 data to stable 2010-era census tract boundaries is seriously limited by reaching only a subset of records.

Table 4 Percent of 2010 census tracts assigned by MAF Match process that existed in contemporaneous years and percent of records in those census tracts, by year

1040 Year	% of assigned 2010 census tracts that existed in 1970	% of records in 2010 census tracts that existed in 1970	% of assigned 2010 census tracts that existed in 1980	% of records in 2010 census tracts that existed in 1980	% of assigned 2010 census tracts that existed in 1990	% of records in 2010 census tracts that existed in 1990
1969	30%	51%				
1974	29%	46%	40%	59%		
1979			39%	55%		
1984			39%	52%	57%	67%
1989					55%	63%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-019.

As part of the DCDL project, we sought to assign small geographic areas to the historical IRS Form 1040 data. This would allow future data users to analyze the data without address-level PII as well as minimizing the need for future geocoding of these unstructured addresses. The census tracts assigned would be contemporaneous to the data themselves. With this approach, the geographic variables would be

coded using the same standard as other small-geography contextual data published by the Census Bureau following each decennial census.

Methods

The DCDL project team used three methods to assign contemporaneous census tracts to the Form 1040 data. We converted the census tracts assigned in the MAF Match process to contemporaneous census tracts through a crosswalk. Second, we geocoded the street address string fields and converted the modern census tracts to contemporaneous census tracts. Finally, we spatially joined the latitudes and longitudes assigned in the MAF Match process to shapefiles of historical census tract maps and assigned census tracts based on the maps. The following section describes each of these in greater detail. All computing took place in a Linux environment (Red Hat, Inc., 2018).

MAF Match-based 2010 census tract crosswalk conversion

The first method converts the 2010 census tract variables from the geographic supplement files to contemporaneous census tracts, using three crosswalks. We developed these crosswalks based on the Longitudinal Tract Database's (Logan et al., 2020) 1970-, 1980-, and 1990-to-2010 census tract crosswalks. The original crosswalks are many-to-many tables with a weighting variable. These tables are intended to be used to translate census tract-level data from one of those three years to 2010 census tracts. Our team did the reverse of that and translated the 2010 census tract assignments to one of the three older years' census tracts. Using the original tables, we created three many-to-one tables using a Python script (Van Rossum & Drake, 2009). This script assigns one 1970, 1980, or 1990 census tract per 2010 census tract, based on the largest weight relative to all contemporaneous census tracts associated with the 2010 census tract in the base file. It also eliminates any census tract relationship which has a weight of zero or that does not fall in the same state or county.

The crosswalks were applied to each 2010 census tract from the geographic supplement files to attempt to identify what census tract the record would have been in at the time the IRS Form 1040 was submitted (1974 and 1984 were converted to both of the two nearest years). Looking back at Table 3, between 71% and 80% of records were assigned 2010 census tracts in the MAF Match process. Approximately 60% of 1969 records were successfully assigned a 1970 census tract—a number that grows over time with 79% of 1989 records converted to a 1990 census tract. Table 5 presents the results of this crosswalk conversion.

Table 5 2010 census tract-assigned records converted to contemporaneous census tracts, by year

1040 Year	Records converted to 1970 census tracts		Records converted to 1980 census tracts		Records converted to 1990 census tracts	
	% of total	% of census tract-assigned	% of total	% of census tract-assigned	% of total	% of census tract-assigned
1969	60%	85%				
1974	61%	85%	65%	90%		
1979			67%	90%		
1984			68%	89%	75%	99%
1989					79%	99%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

Geocoding with SAS and crosswalk conversion

Using SAS software's geocode procedure (SAS Institute Inc., 2013) and maps provided on SAS Institute's (2009) online maps resources, the addresses listed in the IRS Form 1040 main files were geocoded to 2009 TIGER/Line census tracts. Through this process, we were able to geocode between 56% and 64% of records and assign 2009 census tracts (see Table 6). Comparing the results in Tables 3 and 6, the SAS-based geocoding process produced fewer matches than the results in the geographic supplement. This difference makes sense because Census Bureau staff have access to more robust, customizable software than the off-the-shelf version of the procedure we used.

Table 6 Percent of records geocoded using SAS proc geocode process, by year

1040 Year	Percent of records geocoded and assigned census tracts through SAS
1969	60%
1974	56%
1979	64%
1984	58%
1989	64%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

The 1970-, 1980-, and 1990-to-2010 census tract crosswalks described above were applied to the SAS-geocoded 2009 census tracts, attempting to convert the more recent census tract assignments to census tracts contemporaneous to the IRS Form 1040 tax records. Between 40% and 50% of tax records were assigned and converted to a contemporaneous census tract (71-81% of geocoded records could be converted through the crosswalk). Table 7 presents the results of this conversion.

Table 7 Geocoded records converted to contemporaneous census tracts, by year

1040 Year	Records converted to 1970 census tracts		Records converted to 1980 census tracts		Records converted to 1990 census tracts	
	% of total	% of geocoded	% of total	% of geocoded	% of total	% of geocoded
1969	44%	72%				
1974	40%	71%	43%	76%		
1979			47%	75%		
1984			43%	74%	47%	81%
1989					50%	78%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

Spatial joining latitude/longitude to contemporaneous NHGIS TIGER/Line files

The final method of assigning contemporaneous census tracts involves overlaying point-level data from the IRS Form 1040 records onto year-specific census tract boundary shapefiles. The 1970, 1980, and 1990 census tract boundary line files based on the 2000 TIGER/Line files were downloaded from the IPUMS NHGIS program (Manson et al., 2021). The goal of this method is to plot the addresses from the IRS Form 1040 data onto the same plane as the census tract boundaries and append the census tract numbers from those shape files onto the IRS Form 1040 data.

We could not identify a geocoding program that can be run on our available software without accessing an online API, aside from geocode procedure we used (SAS Institute Inc., 2013). Therefore, due to the prohibition of Internet use in the Census Bureau's restricted data environment, we could only use records already assigned to a coordinate plane. This method, therefore, relies on the latitude and longitude data from the geographic supplement files. Between 62% and 73% of records were assigned a latitude and longitude during the MAF Match process (see Table 3).

The spatial join was executed by first ensuring that both the shapefiles and the points used the same projection, requiring reprojection in QGIS (QGIS.org, 2012). Then because the version of QGIS used could not open the Python module, the spatial joining was done outside of QGIS in Python (Van Rossum & Drake, 2009) using the GeoPandas library (Jordahl et al., 2021). Given the size of the data files, the Python code read the data in chunks, ran the spatial join, and output the results to a CSV file. Between 48% and 73% of IRS Form 1040 records were assigned a contemporaneous census tract through this spatial join process. Table 8 presents the results of these spatial joins.

Table 8 Latitude/longitude-assigned records spatially joined to contemporaneous census tracts, by year

1040 Year	Records spatially joined to 1970 census tracts		Records spatially joined to 1980 census tracts		Records spatially joined to 1990 census tracts	
	% of total	% of lat/long-assigned	% of total	% of lat/long-assigned	% of total	% of lat/long-assigned
1969	48%	77%				
1974	55%	86%	58%	91%		
1979			62%	91%		
1984			63%	91%	69%	100% [a]
1989					73%	100% [a]

[a] Over 99%; reported as 100% due to rounding.
All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

Results

Overall, this combination of methods was quite successful at assigning contemporaneous census tracts to the IRS Form 1040 records, and the success at census tract assignment increased over time from 62% in 1969 to 81% in 1989 (see Table 9). This trend is perhaps intuitive, given that much of the United States

Table 9 Summary of records assigned contemporaneous census tracts, by year

1040 Year	Percent of records assigned a 1970 census tract through at least one method	Percent of records assigned a 1980 census tract through at least one method	Percent of records assigned a 1990 census tract through at least one method
1969	62%		
1974	63%	67%	
1979		69%	
1984		69%	77%
1989			81%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

was not divided into census tracts in 1970, while the entire country was divided into census tracts by 1990 (see Figure 2). It may also be assumed that the 1989 maps and street names more closely align with the 21st-century data files used for geocoding than those of 1969. Overall, the conversion of the MAF-assigned 2010 census tracts via a crosswalk assigned census tracts to the most records.

Comparison of census tract assignments

Beyond the assignment of census tracts to IRS Form 1040 records by one or more methods, we also examine the quality of matches. Overall, the number of records successfully assigned contemporaneous census tracts through all three methods increased over time from 34% in 1969 to 46% in 1989 (Table 10). Again, we may intuit that this is due to the increase in the geographic coverage of census tracts and the recency of the later addresses relative to the reference geodata. Looking within those records assigned census tracts through all three methods, the percent of records assigned the same census tract with each method is very high and remains quite steady over time at 91-94%. Therefore, while not every census tract assignment method worked for every record, the three methods seem to have attained consistent matches the vast majority of the time.

Table 10 Summary of assigned contemporaneous census tract consistency, by year

1040 Year	1970		1980		1990	
	Percent of total records assigned census tracts through three methods	Percent of total records assigned same census tract through all three methods	Percent of total records assigned census tracts through three methods	Percent of total records assigned same census tract through all three methods	Percent of total records assigned census tracts through three methods	Percent of total records assigned same census tract through all three methods
1969	34%	31%				
1974	36%	32%	38%	35%		
1979			43%	40%		
1984			40%	37%	43%	40%
1989					46%	43%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

Census tract-divided counties vs. census tract-assigned records

In addition to checking internal consistency, we also assessed whether records that were eligible to be assigned to a census tract were actually assigned to one. We used an IRS Form 1040 records' county variables to determine whether the record fell into a county that was census tract-divided during the filing time (and therefore should have a contemporaneous census tract assigned to it). Using the Longitudinal Tract Database (Logan et al., 2020) crosswalks, we created lists of counties that had at least one census tract in 1970, 1980, and 1990 and merged it with the IRS Form 1040 data files, adding a binary variable indicating whether the record fell in a county with at least one census tract.

Due to the changes in the proportion of the United States that was divided into census tracts, the percent of tax records in a census tract-divided county increases over time, with 76% of records falling in counties that are at least partially census tract-divided in 1969 and 99% of records falling in census tract-divided counties in 1989 (see Table 11). The 1989 number falls short of 100% (the entire country was divided into census tracts in 1990) due to tax records with states coded as foreign state or uncoded.

Table 11 Percent of records in census tract-divided counties and percent of records assigned a contemporaneous census tract variable through at least one method, by year

1040 Year	1970		1980		1990	
	Percent of records in census tract-divided counties	Percent of records assigned a census tract variable through at least one method	Percent of records in census tract-divided counties	Percent of records assigned a census tract variable through at least one method	Percent of records in census tract-divided counties	Percent of records assigned a census tract variable through at least one method
1969	76%	62%				
1974	76%	63%	82%	67%		
1979			81%	69%		
1984			82%	69%	99%	77%
1989					99%	81%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

It should be noted that this method is just used as a check because while most counties would have either been completely census tract-divided or not census tract-divided at all in 1970, about 7% of 1970 census tracts fell outside of Standard Metropolitan Statistical Areas (U.S. Bureau of the Census, 1976). Some counties included only one or a few census tracts; examples of these counties include:

- Story County, Iowa was not fully census tract-divided, while Ames, Iowa was census tract-divided.
- Mower County, Minnesota was not fully census tract-divided, while Austin, Minnesota was census tract-divided.
- Missoula County, Montana was not fully census tract-divided, while Missoula, Montana was census tract-divided.
- Washington County, Virginia was not fully census tract-divided, while areas adjacent to Bristol, Virginia were census tract-divided.
- Wilson County, North Carolina was not fully census tract-divided, while Wilson, North Carolina was census tract-divided.

The percentages in Table 11 represent the records overall, rather than the overlap between census tract-divided counties and census tract assignment for specific records. Comparing on a record level, almost 47 million 1969 records from census tract-divided counties were assigned a 1970 census tract (82% of census tract-divided counties, 62% overall—see Table 12). This changes very little over time, ranging from 77% to 84% of records in census tract-divided counties being assigned a contemporaneous census tract.

The comparison of the remaining records (i.e., those that were not from census tract-divided counties and assigned a contemporaneous census tract) is informative. Looking at Table 12, the percent of records in a non-census-tract-divided county and without a census tract assigned decreases from 24% in 1969 to about 1% in 1989, which is to be expected, as the proportion of the country assigned census tracts increased over time. The number of records in a non-census-tract-divided county but with a census tract assigned is also very low (less than 1% overall). The final category (records in a census tract-divided county without a census tract assigned) accounts for 16% of records overall (13-23% in each year). These unassigned records are likely caused by a combination of addresses that are difficult/impossible to geocode, street maps/addresses that have changed from collection year to the creation of the MAF, and counties that weren't fully census tract-divided.

Table 12 Comparison of census tract-divided counties to census tract-assigned records, by year

1040 Year	Census tract year	County divided into census tracts, census tract assigned through one or more methods	County not divided into census tracts, no census tract assigned	County divided into census tracts, no census tract assigned	County not divided into census tracts, census tract assigned through one or more methods
1969	1970	62%	24%	14%	0%[a]
1974	1970	63%	24%	13%	0%[a]
	1980	67%	18%	15%	0%[a]
1979	1980	69%	19%	13%	0%[a]
1984	1980	69%	18%	13%	0%[a]
	1990	77%	1%	23%	0%[a]
1989	1990	81%	1%	18%	0%[a]

[a] Less than 1%; reported as 0% due to rounding.
All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

Recommendations for using assigned census tract variables

The previous section describes how contemporaneous census tracts were assigned to historical IRS 1040 data in order to ease future analysis, linkage, and aggregation. The following section provides recommendations to researchers on how to use those assigned census tracts.

The project team created an additional census tract variable—recommended census tract—to present future users with a simple way of assessing which census tract value to use. The recommended census tract variable is populated if:

- All three methods assigned the same contemporaneous census tract,
- Two methods assigned the same census tract while the third method did not assign a census tract, or
- Only one method assigned a contemporaneous census tract, while the other two methods did not assign a census tract.

Two-thirds of all records have a recommended contemporaneous census tract assigned, ranging from 58% for 1969 records to 76% for 1989 records. Table 13 presents the percent of records with recommended census tracts assigned. As a point of reference, Table 9 shows that 62% of 1969 records and 81% of those in 1989 had any census tract assigned at all. We also explored recommending census tracts for records assigned the same contemporaneous census tract through two methods and a different census tract through the third method, but that process only increased the percent recommended by about 3% and was determined to be less useful.

Table 13 Percent of records with recommended census tracts, by year

1040 Year	Percent of records with recommended 1970 census tracts	Percent of records with recommended 1980 census tracts	Percent of records with recommended 1990 census tracts
1969	58%		
1974	58%	62%	
1979		64%	
1984		65%	72%
1989			76%

All results were approved for release by the U.S. Census Bureau, Data Management System number: P-7506192 and approval number CBDRB-FY22-ERD002-016.

Limitations

While the project team successfully assigned contemporaneous census tracts to the majority of the historical tax records, several limitations must be noted for future researchers using these data.

The first set of limitations relate to the source data. The addresses recorded in the IRS Form 1040 data are tax filer-reported addresses that were subsequently transcribed—either of those two processes could introduce error. As self-reported street addresses, the data are unstructured and messy, with a range of non-street-based entries including post office boxes and institution names, all of which create barriers to geocoding. The large number of state and county variables assigned to these tax datasets also creates difficulties, especially with regard to checking an assigned census tract against the county variable.

Limitations may also be related to the census tract number crosswalk created based on the Longitudinal Tract Database (Logan et al., 2020). As stated before, the intent of this database is to translate older census tract-level data to 2010 census tracts, and our methods use it in the reverse. To Logan's knowledge, no other researchers have used this resource to reverse the crosswalk (personal communication, March 2022). The project team used the Longitudinal Tract Database's (Logan et al., 2020) weighting factor as a determinant of which historical census tract to assign, which means that while the assigned census tract is probabilistically correct, there is a chance that an address actually falls in a different census tract or no census tract at all. Two of the three methods of assigning contemporaneous census tracts use this reverse crosswalk, and therefore only the spatial join method is not impacted by this limitation.

Third, two of the three methods are reliant upon variables assigned through the MAF Match process (the 2010 census tract and the latitude/longitude). The project team places great confidence in the MAF Match process and uses these data without hesitation. However, any errors inadvertently introduced in that process would impact census tracts we assigned through two methods.

Finally, we acknowledge that, while the U.S. Census Bureau (2018) has historically attempted to delineate census tract boundaries to align with neighborhoods, census tracts prioritize statistical comparability and reliability. As such, well-founded critiques have been made of the use of census tracts rather than geographies more meaningful to local populations (e.g. Lee et al., 2008; Matthews et al., 2021). We fully support the research communities' use of alternative geographic units (e.g., Fowler, Lee, & Matthews, 2016; Lee et al., 2008). At the same time, the depth and breadth of available data with historical census tract designations make processes such as those described here essential to data linkage and research using previously-established methods.

Acknowledgements
We are grateful to Carlos Becerra, John Sullivan, Aneesah Williams, and Jennifer Withrow at the U.S. Census Bureau for providing helpful feedback on this paper. This material is based on work supported by grants from the Hewlett Foundation (2018-7191) and the National Science Foundation (BCS-2023639).

References
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

Esri Inc. (2019). ArcGIS Pro (Version 2.4.3) [Computer software]. Esri Inc. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association, 74*(366a), 269–277. https://doi.org/10.1080/01621459.1979.10482505

Fowler, C. S., Lee, B. A., & Matthews, S. A. (2016). The Contributions of Places to Metropolitan Ethnoracial Diversity and Segregation: Decomposing Change Across Space and Time. *Demography*, *53*(6), 1955–1977. http://www.jstor.org/stable/44161222exit

Genadek, K. R., & Alexander, J. T. (2019). *The Decennial Census Digitization and Linkage Project* (No. 2019–01; ADEP Working Paper Series). U.S. Census Bureau.

Genadek, K. R., & Alexander, J. T. (2022). The missing link: Data capture technology and the making of a longitudinal U.S. Census Infrastructure. *IEEE Annals of the History of Computing*, *44*(4), 57–66. https://doi.org/10.1109/mahc.2022.3195001

Jaro, M. A. (1978). *Unimatch: A record linkage system: Users manual*. Washington, DC: U.S. Bureau of the Census.

Jordahl, K., Van den Bossche, J., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., … Leblanc, F.. (2021, February 28). geopandas/geopandas: v0.9.0 (Version v0.9.0). Zenodo. http://doi.org/10.5281/zenodo.3946761

Lamas, E.J., & Johnson, R.V. (2008, March 10). Transfer of IRS 1040 Historical Files for Tax Year (TY) 1969, TY1974, TY1979 and TY1984 [Unpublished Memorandum]. U.S. Census Bureau.

Lee, B. A., Reardon, S. F., Firebaugh, G., Farrell, C. R., Matthews, S. A., & O'Sullivan, D. (2008). Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, *73*(5), 766–791. https://doi.org/10.1177/000312240807300504

Logan, J. R., Xu, Z., Stults, B. J., & Zhang, C. (2020). *Census geography: Bridging data for census tracts across time.* Brown University. https://s4.ad.brown.edu/Projects/Diversity/Researcher/Bridging.htm

Manson, S., Schroeder, J., Van Riper, D., Kugler, T., & Ruggles, S. (2021). *IPUMS National Historical Geographic Information System: Version 16.0* [dataset]. Minneapolis, MN: IPUMS. 2021. http://doi.org/10.18128/D050.V16.0

Matthews, S. A., Stiberman, L., Raymer, J., Yang, T.-C., Gayawan, E., Saita, S., Tun, S. T. T., Parker, D. M., Balk, D., Leyk, S., Montgomery, M., Curtis, K. J., & Wong, D. W. S. (2021). Looking Back, Looking Forward: Progress and Prospect for Spatial Demography. *Spatial Demography, 9*(1), 1–29. https://doi.org/10.1007/s40980-021-00084-9

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

Perlmann, J. (1979). Using Census Districts in Analysis, Record Linkage, and Sampling. *The Journal of Interdisciplinary History*, *10*(2), 279–289. https://doi.org/10.2307/203338

QGIS.org. (2012). QGIS 1.8.0-Lisboa Geographic Information System [Computer software]. QGIS Association. http://www.qgis.org

Raymer, J., Bai, X., Liu, N., & Wilson, T. (2020). Reconciliation of Australian demographic data to study immigrant population change across space and Time. *Spatial Demography*, *8*(2), 123–153. https://doi.org/10.1007/s40980-020-00060-9

Red Hat, Inc. (2018). Red Hat Enterprise Linux Server release 6.10 (Santiago) [Operating System]. Raleigh, NC: Red Hat, Inc.

SAS Institute, Inc. (2009). 2009 Street Lookup Data for 9.4 [Dataset]. Cary, NC: SAS Institute Inc. https://support.sas.com/downloads/package.htm?pid=2557#

SAS Institute, Inc. (2013). SAS for Unix (Version 9.4) [Computer software]. Cary, NC: SAS Institute Inc.

Spencer, B.D. (1980). Data Used in General Revenue Sharing. In: Benefit-Cost Analysis of Data Used to Allocate Funds. Lecture Notes in Statistics, vol 3. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-6099-8_4

U.S. Bureau of the Census (1972a). *Census of Population and Housing: 1970 Census Tracts Final Report PHC(1)-11 Ann Arbor, Mich. SMSA.* Washington, D.C.: U.S. Government Printing Office. https://www2.census.gov/library/publications/decennial/1970/phc-1/39204513p1ch12.pdf

U.S. Bureau of the Census (1972b). *Census of Population and Housing: 1970 Census Tracts Final Report PHC(1)-226 Washington, D.C.-Md.-Va. SMSA.* Washington, D.C.: U.S. Government Printing Office. https://www2.census.gov/library/publications/decennial/1970/phc-1/39204513p22ch11.pdf

U.S. Bureau of the Census (1976). *U.S. Census of Population and Housing: 1970 Procedural History PHC(R)-1.* Washington, D.C.: U.S. Government Printing Office. http://www2.census.gov/prod2/decennial/documents/1970/proceduralHistory/1970proceduralhistory.zip

U.S. Bureau of the Census (1995). *1990 Census of Population and Housing: History (1990 CPH-R-2C).* Washington, D.C.: U.S. Government Printing Office. https://www2.census.gov/prod2/cen1990/cph-r/cph-r-2.pdf

U.S. Bureau of the Census (2018, October 30). Census Tracts for the 2020 Census—Final Criteria [Docket Number 180927898–8898–01], 83 Fed. Reg. 56277. https://www.govinfo.gov/content/pkg/FR-2018-11-13/pdf/2018-24567.pdf

U.S. Census Bureau. (2000a). 1990 Census tract/BNA outline map (recreated). Washington, D.C.: Census Bureau. https://www2.census.gov/geo/maps/trt1990/st24_Maryland/24033_PrinceGeorges/90T24033_003.pdf

U.S. Census Bureau. (2000b). Census outline map (Census 2000). Washington, D.C.: Census Bureau. https://www2.census.gov/plmap/pl_trt/st24_Maryland/c24033_PrinceGeorges/CT24033_003.pdf

U.S. Census Bureau. (2020). *2020 Census Redistricting Data (Public Law 94-171): P1*. https://data.census.gov/cedsci/table?q=population&t=Populations%20and%20People&g=1400000US24033802405,26161400500&y=2020&tid=DECENNIALPL2020.P1

U.S. Census Bureau. (2022). 2020 TIGER/Line Shapefiles. Washington, D.C.: Census Bureau. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wagner, D., & Layne, M. (2014). *The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software* (No. 2014–01; CARRA Working Paper Series). U.S. Census Bureau.