The Census Historical Environmental Impacts Frame

by

Jennifer R. Withrow
U.S. Census Bureau

Kendall A. Houghton
U.S. Census Bureau

Eva Lyubich
U.S. Census Bureau

Mary Munro
MITRE Corp.

Suvy Qin
U.S. Census Bureau

John L. Voorheis
U.S. Census Bureau

CES 24-66 **October 2024**

Abstract

The Census Bureau's Environmental Impacts Frame (EIF) is a microdata infrastructure that combines individual-level information on residence, demographics, and economic characteristics with environmental amenities and hazards from 1999 through the present day. To better understand the long-run consequences and intergenerational effects of exposure to a changing environment, we expand the EIF by extending it backward to 1940. The Historical Environmental Impacts Frame (HEIF) combines the Census Bureau's historical administrative data, publicly available 1940 address information from the 1940 Decennial Census, and historical environmental data. This paper discusses the creation of the HEIF as well as the unique challenges that arise with using the Census Bureau's historical administrative data.

* Corresponding Author: Jennifer Withrow, jennifer.withrow@census.gov. Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product (Data Management System (DMS) number: P-7505723, Disclosure Review Board (DRB) approval numbers: CBDRB-FY24-CES010-009, CBDRB-FY25-CES025-003).

1. Introduction

This paper details the construction of the Historical Environmental Impacts Frame (HEIF), a data set linking individuals' residences from 1940 through the 1990s to comprehensive measures of local environmental conditions. The HEIF extends backwards in time the Census Environmental Impacts Frame (EIF, Voorheis et al. 2023), which provides detailed annual geocoded data from 1999 to present on individuals' demographics, residential addresses, and environmental exposures. The HEIF is maintained as a separate data file due to the changing nature of the availability and quality of administrative data and the resulting unique data limitations in the twentieth century. The HEIF currently contains precise latitude and longitude[1] of individuals' residential locations in 1940, 1970, 1975, 1980, 1985, 1990, and 1995-1996, with additional years to be added as Census Bureau efforts to digitize and link historical Decennial Census data from 1950-1990 are completed. Historical environmental data, such as historic data on drought, severe weather disasters, and proximity to toxic waste sites can be mapped onto residential location in the HEIF. Taken together with the core EIF, the HEIF makes it possible to follow individuals and their exposures to environmental amenities and hazards over the span of 84 years, creating a new opportunity to better understand the long term, intergenerational, and changing nature of the relationship between environmental conditions and the people of the United States.

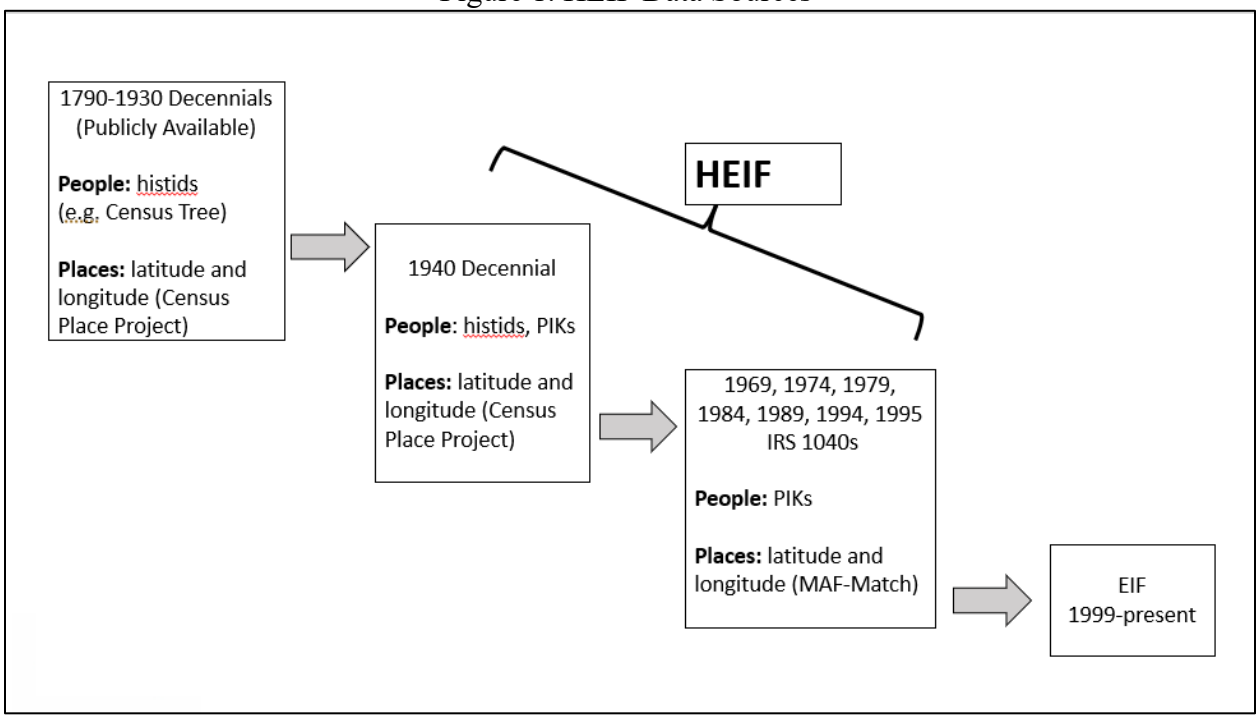
In this paper, we detail the construction of the historical residential history file. We first describe our data sources: the 1940 Decennial Census, the Census Place Project, and IRS 1040 tax returns for the years 1970-1996. We then assess the HEIF's population coverage by comparing the demographic characteristics of our sample to Decennial Census estimates, evaluating the extent to which our data represents the broader US population. While we find that the HEIF, like the EIF, suffers from biases in coverage as not all populations are equally well represented in the underlying administrative records sources, overall we find that our HEIF populations look broadly similar to the U.S. population across most sociodemographic indicators. Finally, we discuss advantages and disadvantages of using the HEIF and the EIF to measure migration. We find that while the HEIF and EIF combined can be a powerful tool for understanding individual migration patterns over a long period, using the HEIF and EIF to understand country-level patterns in migration propensities presents some challenges due to fundamental differences between survey and administrative definitions and reference periods of migration.

2. Constructing the Historical Residential History File (RHF)

Figure 1 lists the data sources for each year of the residential history portion of the HEIF. In the following sections, we describe each data source in detail. Years 1940 through 1996 of the HEIF are available under select approved projects inside the Census Bureau's secure computing environment.

[1] Latitude and longitudes allow users to assign different levels and/or vintages of Census or administrative geography, depending on their environmental or other data of interest.

Figure 1. HEIF Data Sources



Notes: Figure shows the main data sources and linkage variables between the Historical Environmental Impacts Frame, the Environmental Impacts Frame, and publicly available residential history sources.

1940 Decennial Census

1940 residential latitudes and longitudes come from The Census Place Project (Berkes, Karger, and Nencka 2023), an effort to geocode sub-county address data from the 1790-1940 decennial censuses. These places are assigned a latitude and a longitude, as well as county and state FIPS codes based on 2016 Census Bureau cartographic boundaries. We link Census Place Project data with the Census Bureau's internal version of the 1940 Decennial via histids, or historical census identification numbers[2] to obtain Protected Identification Keys (PIKs) created internally by the Person Identification Validation System (PVS) (Wagner and Layne 2014), which allow the 1940 Census to be linked with other data sources within the Census Bureau. 1940 histids are maintained, allowing researchers expand residential histories by using publicly available crosswalks between the 1940 and earlier decennial censuses, such as the Census Linking Project and the Census Tree, to make further linkages backward as far as 1790 (Abramitzky et al. 2020; Price et al. 2021; Buckles et al. 2023).

1969-1995 1040 Tax Returns

Our main source of address data for the 1970-1996 period come from digitized versions of Internal Revenue Service (IRS) 1040s held at the Census Bureau. IRS 1040s are available for tax years 1969, 1974, 1979, 1984, 1989, 1994 and 1995. Alexander et. al. (2024) and Genadek et. al. (2024) discuss these files in depth, including the various geographic variables available and coverage based on IRS taxpayer estimates. Digitized 1969 1040s only include primary filers, 1974-1989 1040s include the primary and secondary filers, and 1994 and 1995 1040s include

[2] Histids are identification numbers assigned to the public use version of IPUMS complete count historical census data.

both primary and secondary filers as well as up to four dependents. Starting in tax year 1979, filing dates are included. In the case that there is more than one entry per individual, we follow the same de-duplication process as in Voorheis et al. (2023).[3] Consistent with the core EIF, we use processing year timing for the 1040 files, so will refer to locations in the 1969 tax year 1040s in terms of the year in which they were filed (1970), as this timing more closely corresponds to residence at tax filing.

When the Census discovered and digitized these IRS records, they also conducted a process to assign specific addresses from the Census Bureau's Master Address File (MAFIDs), and census tracts to the address information available based on the 2017 MAF and 2010-era MAFIDs (Bleckley, Genadek, and Alexander 2023; Onorato and Winkelmann 2018; Wagner and Layne 2014). In this version of the HEIF, we keep those individuals who were successfully matched to a MAFID, which allows us to assign a specific latitude and longitude. A large portion of street addresses were not able to be assigned a MAFID. The 1969 1040s have the lowest coverage, with 62 percent of tax records assigned a MAFID, rising to over 75 percent in the 1990s (Onorato and Winkelmann 2018). Future iterations of the HEIF will leverage additional address information available at the tract and county level in the original 1040 files to expand coverage (see Bleckley, Genadek, and Alexander (2023)).

Demographic Spine

The HEIF uses the Census Bureau Numident as its source of demographic information, as in the EIF. The Numident provides information on date of birth, date of death, and place of birth for all individuals who have ever applied for a Social Security Number. Additionally, race and ethnicity information is included using an internal Census Bureau file that harmonizes race information from surveys and administrative records.

Future Data sources

As the digitization of historic administrative and survey data continues, the HEIF will be able to cover a more continuous set of years. For example, the upcoming digitization and assigning of PIKs to the 1950 and 1990 decennial censuses will provide important residential history information for our intervening years.

3. Evaluating Coverage and Data Quality

While administrative records linked through the Census Bureau's PVS program provide information on large numbers of US residents, certain groups are excluded when relying only on administrative records. First, as demographic information comes from the Census Numident, we only include individuals that have applied for a Social Security Number or Individual Taxpayer Identification Number (ITIN).[4] Second, certain groups are more likely to receive a PIK than others -- such as those who are non-Hispanic White and those with a higher income -- due to their higher likelihood of being found in administrative records. Third and importantly for the HEIF compared to the EIF, our only available data source for residential histories from 1970

[3] We first select the entry where the PIK is the primary filer. If the PIK never appears as a primary filer, we select their secondary filer entry, followed by dependent entry if secondary is not available. If multiple entries still exist for an individual, we then select the latest filed return. Any remaining duplicates are removed via random number.
[4] Individual Taxpayer Identification Numbers (ITINs) are not available prior to 1996.

through 1996 are the IRS 1040s, meaning we only observe individuals who have a formal connection to the economy and are 1040 tax filers. Furthermore, only primary filers are recorded in tax year 1969 and only primary and secondary filers are available for tax years 1974, 1979, 1984, and 1989. Tax years 1994 and 1995 include both primary and secondary filers, as well as up to four dependents.

To assess coverage and representativeness, we compare the HEIF with our demographic spine, the Decennial Censuses for years 1970, 1980, and 1990, as well as with the Current Population Survey's Annual Social and Economic Supplement (CPS-ASEC) from 1979, 1984, and 1995. While each dataset has its own drawbacks in terms of suitability of comparison, these differences in suitability help to highlight the special characteristics of the population covered by a residential history file based primarily on tax return data.

1940 PIKs

As part of a large effort by the Census Bureau to expand its linkable data back in time, the digitized 1940 Decennial Census was put through the PVS process to assign PIKs to individual respondents. The PVS process relies on names, age, addresses, presence of a SSN, and other demographic information to assign a unique individual identifier. The availability of this information can vary, leading to the linked population looking slightly different than the full population. In the case of 1940, PIKs were more likely to be assigned to White and native-born respondents, which mirrors the biases in the linking of other administration and survey records. Additionally, there is a lower PIK rate for older individuals in 1940, highlighting the importance of having an SSN (Massey et al. 2018).

Demographic Spine

Comparing our Historical EIF to the demographic spine presents some drawbacks also present in the core EIF: we cannot fully account for the size of the emigrant and deceased populations. Additionally, the population of tax filers is fundamentally different than the population of those who applied to SSNs, who may or may not have left the country or passed away. Additionally, for tax year 1969 through 1989 we do not have information on dependents, meaning we see far fewer children than in the spine. Table 1 shows the average characteristics of those in the HEIF compared to the Spine.

Table 1. Sample Demographics

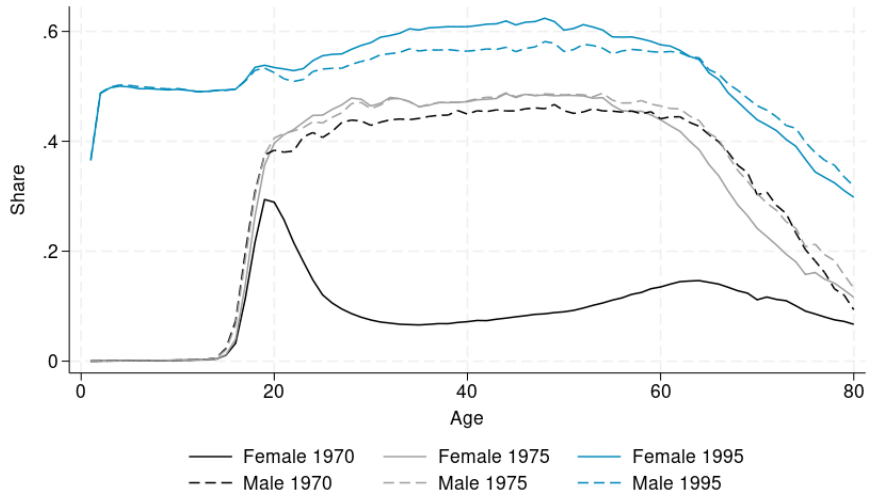
	1940		1970		1980		1990	
	Spine	Spine + RHF	Spine	Spine + RHF	Spine	Spine + RHF	Spine	Spine + RHF
Share Female	0.50	0.50	0.50	0.23	0.50	0.50	0.50	0.50
Mean Age	26.29	21.47	32.84	40.79	35.03	40.68	36.92	42.67
Share <18	0.38	0.47	0.35	0.03	0.29	0.02	0.27	0.02
Share 65+	0.02	0.01	0.12	0.10	0.14	0.10	0.16	0.13
Share Hispanic	0.04	0.01	0.09	0.04	0.10	0.05	0.12	0.07
Share NH White	0.81	0.89	0.73	0.86	0.71	0.83	0.69	0.80
Share NH Black	0.12	0.09	0.12	0.09	0.13	0.09	0.13	0.09
Share NH Asian	0.012	0.001	0.03	0.01	0.03	0.01	0.04	0.02
N	153,000,000	57,630,000	273,100,000	40,660,000	302,800,000	84,010,000	333,700,000	101,000,000

Source: Historical Environmental Impacts Frame Spine and Residential History File, 1940-1990. *Notes:* See Section 2 for details on construction. Table shows characteristics of non-deceased individuals in the spine and non-deceased individuals with an address on the residential history file.

Looking at 1940, our sample is younger and has a greater share of non-Hispanic White individuals than the general resident U.S. population due to the PVS process (Massey et al. 2018). Notably after 1940, when our data source is the IRS 1040s, our sample is older, particularly prior to 1995 as only primary and secondary filers are included. Again, there is a higher share of non-Hispanic White individuals in the residential history file (RHF) compared to the Spine. In 1970, the inclusion of only primary filers in the IRS 1040s leads to a much lower share of women in the residential history file. This share is corrected after 1970 once secondary filers are included.

Figure 2 shows how the coverage of women improves after 1970. We also see a sharp increase in coverage for all at ages 18 to 19 (the age a non-full-time student could no longer be claimed as a dependent on a 1040) with a decline in coverage after retirement age.

Figure 2: Share of HEIF Individuals in Spine, by Year and Age



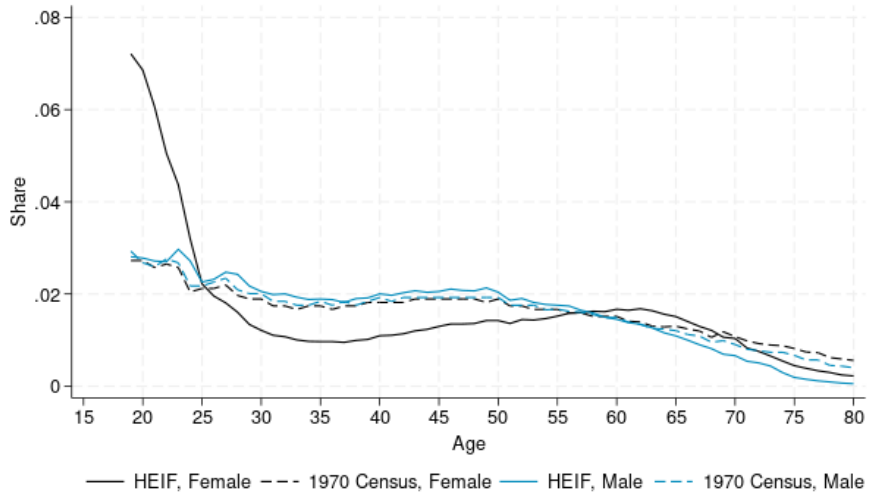
Source: Historical Environmental Impacts Frame Spine and Residential History File, 1970, 1975, and 1995. *Notes:* See Section 2 for details on construction. This figure shows the share of individuals in the spine also present in each year of the HEIF residential history file, by age.

Tax Filers, 1969-1995 and Decennial Censuses

Given that our only source of residential history data for 1970 through 1996 are 1040 returns, we also compare the HEIF population to the universe of tax filers for those years based on the 1970, 1980, and 1990 decennial censuses. The universe of tax filers in each tax year is influenced by, and changes over time due to, tax law. Factors such as the creation of the Earned Income Tax Credit in 1975, changes in income filing thresholds that vary by age, changes in age limits or minimums for dependents and retirees, and changes to which individuals have been digitized from the forms all affect who is present in our data. Unlike our Numident-based demographic spine which will sometimes have incomplete death information for those born in earlier years, we know that respondents to the 1970, 1980, and 1990 Decennial Censuses were alive and residents of the United States on Census day (April 1). We compare HEIF to Census-sourced distributions of age, race and gender for a population aged 19 to 80.

Figures 3 and 4 show the age distributions of individuals in the HEIF compared to the 1970-1990 population censuses. Figure 3 breaks down the age distribution by gender to better emphasize the gender disparities caused by only having primary filers in 1970.

Figure 3. 1970 Age Distributions by Gender, HEIF and Decennial Censuses

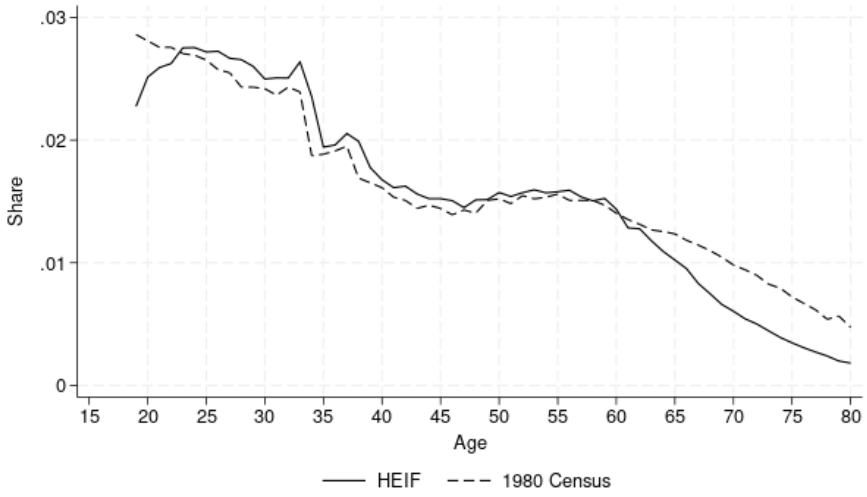


Source: Historical Environmental Impacts Frame Spine and Residential History File, 1970, and 1970 Decennial Census. *Notes:* See Section 2 for details on construction. This figure shows the age distributions by gender in the HEIF and the 1970 Decennial Census.

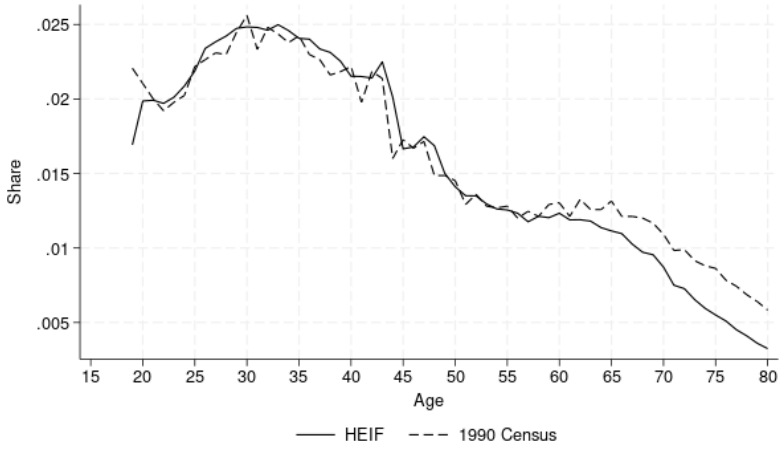
Figure 4 shows the age distributions for men and women combined for both 1980 and 1990. In 1980 we find that the HEIF has lower coverage of the young (under 25) and the old (over 65). In 1990, there is a similar under coverage of the young, but also an overrepresentation of individuals 30 to 50 years old compared to the Decennial Census population.

Figure 4. Age Distributions, HEIF and Decennial Censuses

A. 1980



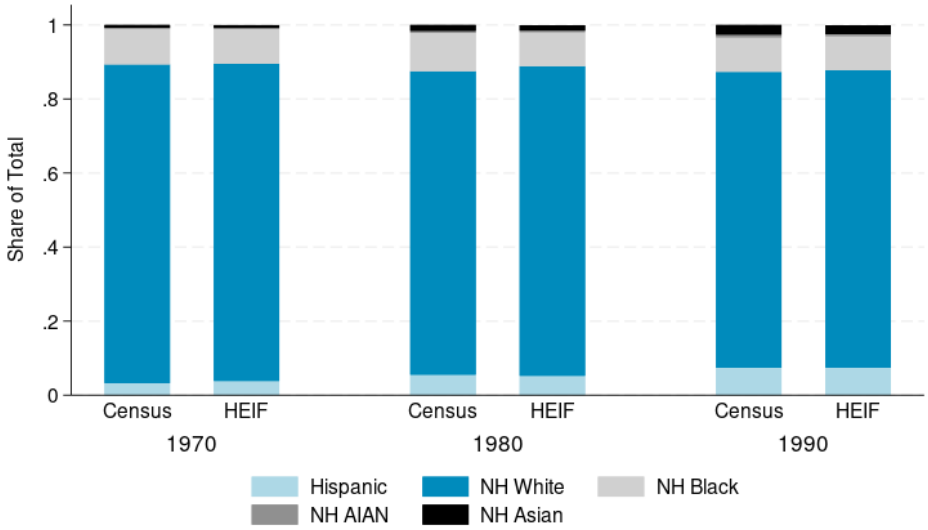
B. 1990



Source: Historical Environmental Impacts Frame Spine and Residential History File, 1980 and 1990, and 1980 and 1990 Decennial Census. *Notes:* See Section 2 for details on construction. This figure shows the age distributions in the HEIF and the Decennial Censuses.

Figure 5 show race distributions between the decennial censuses and the HEIF. Overall, the race distributions are similar across both sources.

Figure 5. Race Distributions, HEIF vs. Decennial Censuses



Source: Historical Environmental Impacts Frame Spine and Residential History File, 1970-1990, and 1970-1990 Decennial Census. *Notes:* See Section 2 for details on construction. This figure shows the distributions by race in the HEIF and the Decennial Censuses.

Finally, Table 2 shows the distribution of state of residence in the HEIF. These states closely correspond[5] to the states with the highest resident populations in each year based on Decennial Census data, meaning tax data are not overrepresenting residents of one state over another compared to the population.

Table 2. Top 5 States by Population

State (Percent of 1040 Filers)													
	1970		1975		1980		1985		1990		1995		1996	
	Share 1040 Filers with a MAFID													
1	CA	12.6	CA	12.9	CA	13.0	CA	12.7	CA	12.9	CA	12.0	CA	11.9
2	NY	10.5	NY	8.2	NY	7.3	NY	6.9	TX	6.6	TX	7.1	TX	7.1
3	PA	6.4	IL	6.1	TX	6.4	TX	6.5	NY	6.1	NY	6.0	NY	6.0
4	IL	6.3	OH	5.9	OH	5.8	OH	5.5	FL	5.6	FL	5.7	FL	5.7
5	OH	6.0	TX	5.8	IL	5.7	IL	5.3	OH	5.2	OH	4.8	IL	4.7

Source: IRS 1040s. *Notes*: Table shows top 5 states of residence in each calendar year by share of 1040 filers who have a MAFID in the 1969-1995 tax year data.

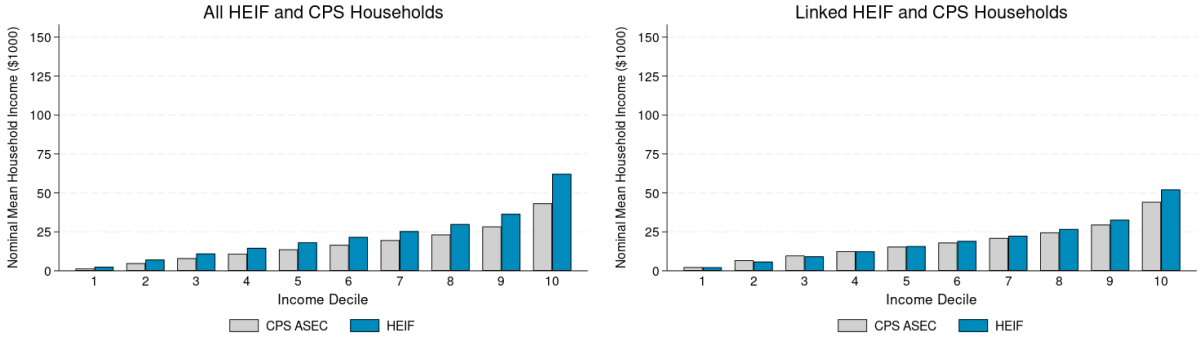
[5] In 1970, the top 5 most populous states in the US were: CA, NY, PA, TX, IL. In 1980: CA, NY, TX, PA, IL. In 1990: CA, NY, TX, FL, PA.

CPS ASEC

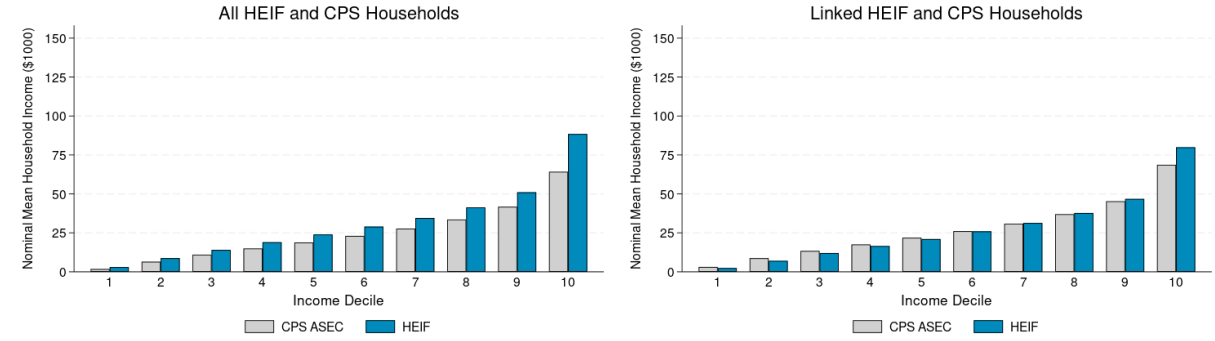
We then compare the adjusted gross income (AGI) of households (summed at the MAFID level) in the HEIF to the household AGI reported in the CPS-ASEC. In Figure 6 we show three HEIF years: 1980, 1985, and 1996. Figures on the lefthand side show the distributions of AGI for all households in each of the datasets, while Figures on the righthand side show the distributions of AGI for households that are present both in the IRS 1040 data and the CPS-ASEC data, linked by PIKs.

Figure 6: Income Distributions, IRS 1040s vs. CPS-ASEC

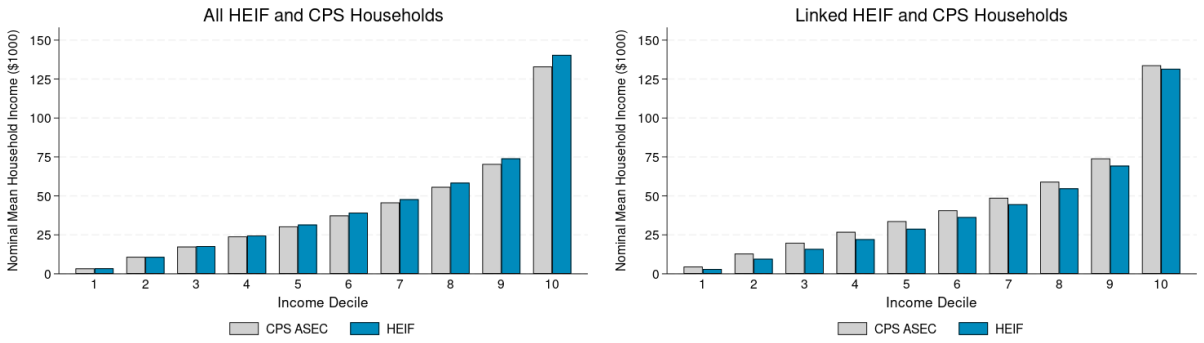
A. 1980 (tax year 1979)



B. 1985 (tax year 1984)



C. 1996 (tax year 1995)



Source: Historical Environmental Impacts Frame Spine and Residential History File, 1980-1996, and CPS-ASEC.
Notes: Household income from IRS 1040s is measured as unique wage and salary, summed at the MAFID level.

Household income in CPS is from the HWSVAL variable. Linked households refer to whether an individual from a household is present in both the IRS 1040s and CPS-ASEC, linked via PIK.

Among all households, we see the household income of those in the HEIF is higher than that reported in the CPS-ASEC. Among those households that have an individual present in both the HEIF and the CPS-ASEC, we see that the pattern does not hold: among some deciles, measured household income in the HEIF is greater than in the CPS-ASEC. This is especially true in 1996. This means that we cannot rule out that the HEIF has under-coverage of low-income households, rather than differences in income instead reflecting different income concepts or survey reporting errors.

As with many historical sources linked over time, availability of data and challenges in linking individuals from marginalized or small populations can lead to underrepresentation. Researchers may choose to address this in part through reweighting (Bailey et al. 2020; Cefalu et al. 2024).

4. Potential Use Cases of the HEIF and Historical Environmental Data

Just as digitization efforts have greatly improved information about individual residential histories, efforts to expand environmental data to the past also provide an opportunity to better measure longitudinal exposure to environmental conditions. The HEIF allows researchers to leverage both developments to build on previous work studying how historical environmental exposures and shocks affect various economic outcomes.

A large literature leverages a diverse set of historical environmental data to study the long-run and dynamic impacts of environmental hazards on various outcomes, such as life expectancy, migration, family formation, agricultural development, and economic growth. For instance, Boustan et al. (2020) use Red Cross disaster relief records to analyze the impact of natural disasters on migration rates, home prices, and poverty rates from 1930 to 2010; Arenberg and Neller (2023) employ fire atlases to investigate the consequences of early-life exposure to smoke between 1930 and 1969; and Kiaghadi, Rifai, and Dawson (2021) explore the effects of toxic emissions inventories on life expectancy. Several studies have focused on specific 20th century environmental crises such as the Dust Bowl (Gutmann et al. 2016; Hornbeck 2012; 2023), Boll Weevil infestations (Bloome, Feigenbaum, and Muller 2017; Lange, Olmstead, and Rhode 2009), the Great Mississippi Flood of 1927 (Hornbeck and Naidu 2014), and the 1950s drought (Rajan and Ramcharan 2023). Others have utilized historical climate data from the National Climatic Data Center (NCDC) Global Historical Climatology Network Daily (GHCN-Daily) to examine how climate shocks affect economic activity, agricultural output, mortality, and adaptation, to name a few (Dell, Jones, and Olken 2014; Schlenker and Roberts 2009; Barreca et al. 2016, among others). Finally, there has been a growing effort to quantity damages from flooding and storms (e.g. Pielke, Jr., Downton, and Barnard Miller 2002; Pielke, Jr. et al. 2008; Raker 2020); Reports on historic flooding events from the U.S. Geological Survey (Historical Flooding | U.S. Geological Survey (usgs.gov) starting in 1900, the Global Flood Database (formerly Dartmouth Flood Observatory), the Atlantic Hurricane Database (HURDAT2) covering every year from 1851-2022, and the National Oceanic and Atmospheric Administration (NOAA) Storm Prediction Center's Severe Weather Database starting in 1950 (Storm Prediction Center Severe Weather GIS (SVRGIS) Page (noaa.gov)) all have potential for integration with the HEIF.

Many of the studies in this literature use measures of environmental exposure at the county level or some other aggregated level, often due to the limited ability to observe outcomes at the individual level--- a constraint that the HEIF relieves. By providing precise residential histories over multiple years, the HEIF enables researchers to narrowly define geographic area of residence subject to exposure and better ensure the individual was living in an exposed area at a given time, and for how long. This enhanced granularity offers the potential to significantly refine our understanding of the economic consequences of environmental conditions.

5. Migration in the HEIF and EIF

While the EIF and HEIF are powerful as cross-sectional datasets, their value is increased by the Census Bureau's PVS linking infrastructure, which makes it possible to use them as longitudinal sources of residential histories. Longitudinal residential histories are particularly important for measuring migration. We provide some basic descriptive statistics of migration based on our HEIF and EIF populations and how those compare to Census Bureau survey measures of migration.[6]

We first construct five-year migration estimates using the HEIF and core EIF from 1970 through 2020. We then compare these estimates to the five-year estimates of migration collected in survey data through the CPS-ASEC and the decennials. Table 3 shows the MAFID, county, and state migration rates for each year from the HEIF and the EIF.

[6]See Sullivan, Genadek, and Bleckley (2023) for a detailed look at migration in the historic tax data (1970-1989) by demographic characteristics, as well as when using all geographic information from these files, not just those assigned a MAFID.

Table 3. Migration Rates, HEIF and EIF

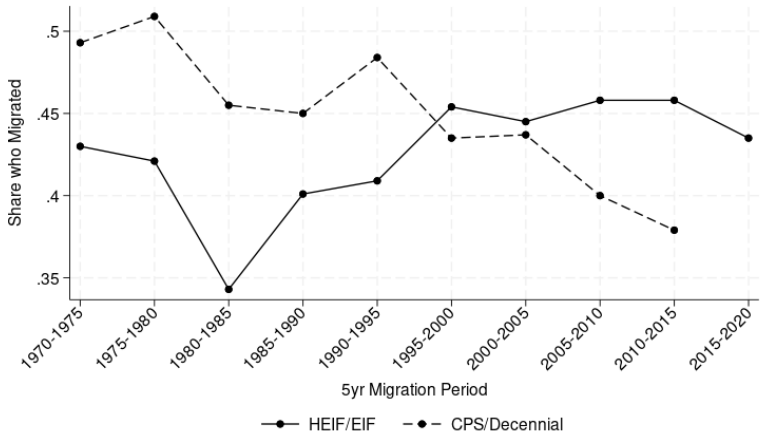
		Type of Move (Fraction of EIF PIKs seen twice)		
5 Year Range	Share of PIKS in year t seen in year t+5	MAFID Movers	County Movers	State Movers
1970 to 1975	0.7087	0.4300	0.1703	0.0866
1975 to 1980	0.7622	0.4213	0.1641	0.0845
1980 to 1985	0.7423	0.3430	0.1396	0.0742
1985 to 1990	0.7690	0.4010	0.1623	0.0819
1990 to 1995	0.8008	0.4095	0.1691	0.0835
1995 to 2000	0.9439	0.4543	0.1862	0.0902
2000 to 2005	0.9081	0.4448	0.1781	0.0831
2005 to 2010	0.8511	0.4582	0.1901	0.0887
2010 to 2015	0.9415	0.4582	0.1926	0.0889
2015 to 2020	0.9514	0.4350	0.1880	0.0882

Source: Historical Environmental Impacts Frame and Environmental Impacts Frame Residential History File, 1970-2020. *Notes:* See Section 2 for details on construction. Tables shows the five-year migration rates for individuals present in year t and t+5.

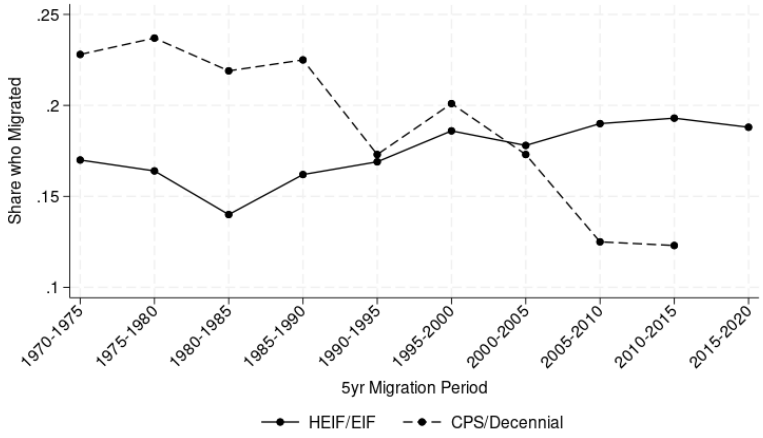
We next compare our HEIF and EIF migration rates with those in other Census surveys: the CPS ASEC and the Decennial Censuses in Figure 7.

Figure 7: Migration Rates, HEIF and EIF vs. Survey Data and Decennial Censuses

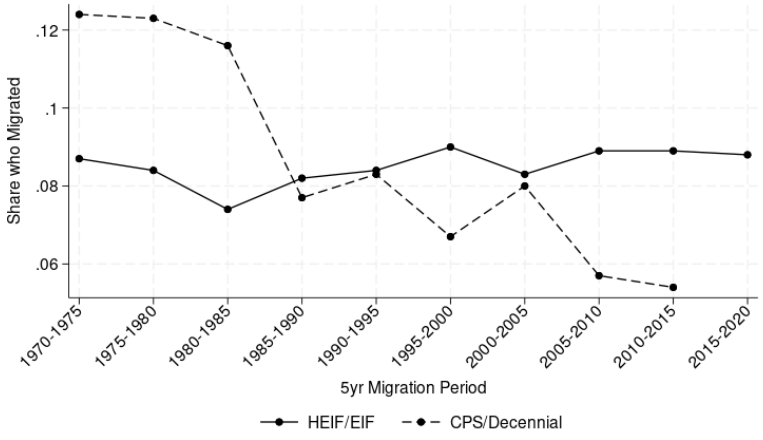
A. Mafid Migrants



B. County Migrants



C. State Migrants



Source: Historical Environmental Impacts Frame and Environmental Impacts Frame Residential History File 1970-2020, CPS-ASEC 1975,1980,1985,1995,2005,2010, and 2015, and Decennial Censuses 1990 and 2000. *Notes:* See Section 2 for details on construction. Tables shows the five-year migration rates for individuals present in year t and t+5 from the Residential History File and 5-year migration rates from the CPS-ASEC and Decennial Censuses.

The discrepancies in migration rates found in the HEIF compared to Census survey data could be the result of several sources, as explored in studies of migration rates of more recent time periods. For example, Ihrke et al. (2015) – when comparing CPS-ASEC and American Community Survey migration estimates -- point to differences in reference periods, question wording, mode of data collection, and timing of data collection as several factors among others that could contribute to disparities between migration estimates. Foster et al. (2018) compare 2010 IRS records to the 2010 ACS and Census microdata and find that, overall, survey respondents tended to underreport migration compared to documented IRS changes of address at the state level. These authors also refer to how fundamental differences in the definition of migration between the ACS and changes of addresses in IRS forms is likely contributing to these discrepancies. Hyatt et al. (2018) also compare CPS, ACS, and Longitudinal Employer-Household Dynamics (LEHD) migration statistics to IRS records and also find discrepancies between the sources.

In the case of the HEIF and EIF, sample selection (especially in the case of the HEIF), timing of administrative records compared to survey responses, migration as measured by a change of address where the individual has financial responsibility compared to a change in where the individual may consider home, are just a few of the factors likely contributing to differences. These comparisons indicate that using the HEIF and EIF residential histories as a measure of overall migration prevalence in the United States presents some important differences with survey data that warrant additional investigation and may be driven in part by compositional differences between the EIF, HEIF and the population. However, using individual measures of migration in the context of response to environmental or other events is still accurate in our setting, as we expect we are capturing legitimate moves.

6. Conclusion and Next Steps

In this paper, we described the creation of the Census Historical Environmental Impacts Frame (HEIF), an extension back in time of the Census Environmental Impacts Frame (EIF). The HEIF is a combination of the residential histories of U.S. residents from 1940-1996 and historical environmental data. It provides a bridge from publicly-available linked Decennial Census data from 1870-1940, the Census Bureau's earliest linkable data to other Census internal data products (1940), and the EIF which begins in 1999.

While the HEIF suffers from some biases in coverage due to both its limited data sources (IRS 1040s for the years 1970-1996) and the PVS process, we show that, in general, the HEIF broadly matches the U.S. population in each year. These residential histories allow researchers to attach fine geographic environmental data at the individual level. In the future, further digitization and linking of Decennial Censuses from 1950 onward will greatly expand our coverage of the U.S. resident population.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Santiago Pérez, and Myera Rashid. 2020. *Census Linking Project* (version 2.0).

Alexander, J. Trent, David A. Bleckley, Jonathan Fisher, Katie Genadek, Susan Hautaniemi Leonard, and Aristotle Magganas. 2024. "Newly Available Individual-Level U.S. Tax Data from 1969-1994." *ADEP Working Paper Series 2024-02*, February.

Arenberg, Samuel, and Seth Neller. 2023. "Ashes to Ashes: The Lifelong Consequences of Early-Life Wildfire Exposure." *Working Paper*. https://sethneller.github.io/papers/Ashes_to_Ashes_Working_Paper.pdf.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data." *Journal of Economic Literature* 48 (4): 997–1044.

Barreca, Alan, Karen Clay, Olivier Deschenes, Michael Greenstone, and Joseph S. Shapiro. 2016. "Adapting to Climate Change: The Remarkable Decline in the US Temperature-Mortality Relationship over the Twentieth Century." *Journal of Political Economy* 124 (1).

Berkes, Enrico, Ezra Karger, and Peter Nencka. 2023. "The Census Place Project: A Method for Geolocating Unstructured Place Names." *Explorations in Economic History* 87:101477.

Bleckley, David A., Katie R. Genadek, and J. Trent Alexander. 2023. "Assigning Contemporaneous Census Tracts to Historical Income Tax Data." *ADEP Working Paper Series 2023-03*.

Bloome, Deirdre, James Feigenbaum, and Christopher Muller. 2017. "Tenancy, Marriage, and the Boll Weevil Infestation, 1892-1930." *Demography* 53 (3): 1029–49.

Boustan, Leah Platt, Matthew E. Kahn, Paul W. Rhode, and Maria Lucia Yanguas. 2020. "The Effect of Natural Disasters on Economic Activity in US Counties: A Century of Data." *Journal of Urban Economics* 118 (July):103257.

Buckles, Kasey, Adrian Haws, Joseph Price, and Haley Wilbert. 2023. "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project." *NBER Working Paper No. 31671*.

Cefalu, Matthew, John Sullivan, Narayan Sastry, Elizabeth Fussell, and Todd Gardner. 2024. "Gradient Boosting to Address Statistical Problems Arising from Non-Linkage of Census Bureau Datasets." *CES Working Paper Number CES-24-27*, June.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2014. "What Do We Learn from the Weather? The New Climate-Economy Literature." *Journal of Economic Literature* 52 (3): 740–98.

Foster, Thomas B., Mark Ellis, and Lee Fiorio. 2018. "The Opportunities and Challenges of Linked IRS Administrative and Census Survey Records in the Study of Migration." *CARRA Working Paper Series Number 2018-06*.

Genadek, Katie, Trent Alexander, Nishaad Rao, and Daniel Chapman. 2024. "How Representative Are Tax Data for Research? Comparing Full Universe Tax Filings Data with U.S. Census Data." *Census Bureau Working Paper*.

Gutmann, Myron P., Daniel Brown, Angela R. Cunningham, James Dykes, Susan Hautaniemi Leonard, Jani Little, Jeremy Mikecz, Paul W. Rhode, Seth Spielman, and Kenneth M. Sylvester. 2016. "Migration in the 1930s: Beyond the Dust Bowl." *Social Science History* 40 (4): 707–40.

Hornbeck, Richard. 2012. "The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe." *American Economic Review* 102 (4): 1477–1507.
———. 2023. "Dust Bowl Migrants: Environmental Refugees and Economic Adaptation." *Journal of Economic History*, July, 1–31.
Hornbeck, Richard, and Suresh Naidu. 2014. "When the Levee Breaks: Black Migration and Economic Development in the American South." *American Economic Review* 104 (3): 963–90.
Hyatt, Henry, Erika McEntarfer, Ken Ueda, and Alexandria Zhang. 2018. "Interstate Migration and Employer-to-Employer Transitions in the U.S.: New Evidence from Administrative Records Data." *CES Working Paper Number CES-16-44R*, May.
Ihrke, David, William Koerber, and Alison Fields. 2015. "Comparison of Migration Data: 2013 American Community Survey and 2013 Annual Social and Economic Supplement of the Current Population Survey." *SEHSD Working Paper Number SEHSD-WP2015-21*.
Kiaghadi, Amin, Hanadi S. Rifai, and Clint N. Dawson. 2021. "The Presence of Superfund Sites as a Determinant of Life Expectancy in the United States." *Nature Communications* 12.
Lange, Fabian, Alan L. Olmstead, and Paul W. Rhode. 2009. "The Impact of the Boll Weevil, 1892-1932." *The Journal of Economic History* 69 (3): 685–718.
Massey, Catherine G., Katie R. Genadek, J. Trent Alexander, Todd K. Gardner, and Amy O'Hara. 2018. "Linking the 1940 Census with Modern Data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51 (4): 246–57.
Onorato, D, and J Winkelmann. 2018. "Assigning Tracts to 1040 Forms." Memorandum. U.S. Census Bureau.
Pielke, Jr., R.A., M.W. Downton, and J.Z. Barnard Miller. 2002. *Flood Damage in the United States, 1926–2000: A Reanalysis of National Weather Service Estimates*. Boulder, CO: UCAR.
Pielke, Jr., R.A., Joel Gratz, Christopher Landsea, Douglas Collins, Mark A. Saunders, and Rade Musulin. 2008. "Normalized Hurricane Damage in the United States: 1900-2005." *Natural Hazards Review* 9 (1).
Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. 2021. "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set." *Explorations in Economic History* 80:101391.
Rajan, Raghuram, and Rodney Ramcharan. 2023. "Finance and Climate Resilience: Evidence from the Long 1950s US Drought." *NBER Working Paper No. 31356*, June.
Raker, Ethan J. 2020. "Natural Hazards, Disasters, and Demographic Change: The Case of Severe Tornadoes in the United States, 1980-2010." *Demography* 57 (2): 653–74.
Schlenker, Wolfram, and Michael J. Roberts. 2009. "Nonlinear Temperature Effects Indicate Severe Damages to US Crop Yields under Climate Change." *Proceedings of the National Academy of Sciences* 106 (37): 15594–98.
Sullivan, John, Katie Genadek, and David Bleckley. 2023. "New Estimates of Geographic Mobility from Individual Level Tax Records (1969-1989)." *Working Paper*.
Voorheis, John, Jonathan Colmer, Kendall Houghton, Eva Lyubich, Mary Munro, Cameron Scalera, and Jennifer Withrow. 2023. "Building the Prototype Census Environmental Impacts Frame." *CES Working Paper Number CES-23-20*.
Wagner, Deborah, and Mary Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA)

Record Linkage Softward." *U.S. Census Bureau Working Paper* No. 2014-01. https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf.