NATIONAL HEALTH AND AGING TRENDS STUDY (NHATS)

1940 CENSUS LINKAGE

USER GUIDE

June 22, 2021

Suggested Citation: Freedman, Vicki A., John Robert Warren, Fabian Pfeffer, Jonas Helgertz, Dafeng Xu, Judith D. Kasper. National Health and Aging Trends Study 1940 Census Linkage User Guide. Baltimore: Johns Hopkins University School of Public Health. Available at www.NHATS.org.

Acknowledgment: This guide was prepared with funding from the National Institute on Aging (R01AG050300 and U01AG032947). The linking received support from the National Institute on Aging (R01AG050300) to Warren. The project also benefitted from administrative and technical support from the Minnesota Population Center, which receives core infrastructure support from NIH grant 2P2CHD041023. Linkages were conducted in the Michigan Center on the Demography of Aging (MiCDA) virtual data enclave, which is funded by the National Institute on Aging P30 AG012846.

INTRODUCTION

This guide describes the linkage of the National Health and Aging Trends Study (NHATS) to the 1940 Census. We briefly describe the NHATS eligible sample and linking rate, the structure of the resulting linked files and how they can be accessed, and linkage procedures including assessments of reliability and validity. An appendix provides a listing of variables and brief descriptions.

The project is part of a larger effort to conduct parallel linkages to the 1940 Census for respondents to NHATS, the Panel Study of Income Dynamics (PSID), the Health and Retirement Study (HRS), the Wisconsin Longitudinal Study (WLS), and the National Social Life, Health, and Aging Project (NSHAP).

NHATS ELIGIBLE SAMPLE AND LINKING RATE

In 2015 (Round 5), NHATS administered the Names Module (NA) to participants who were alive and living in the U.S. on April 1, 1940. The purpose was to obtain information that would improve the chances of linking to their 1940 Census record. The NA module collected name in 1940 and the names and relationships of one or more individuals who they lived with at that time (e.g. a mother and/or father or another person). The Names module (NA) was administered to 4,138 eligible participants.

Of the 4,138 eligible cases, 2,595 (62.7%) were linked to their 1940 Census record. Assessments of the validity and reliability of the linking suggested these matches may be considered high quality.

OVERVIEW OF FILES

NHATS offers users three restricted data files through the NHATS Restricted Data Repository. 1940 Census variable names were retained on these files but data were restructured as follows:

- The NHATS Round 5 1940 Census SP file includes information from the 1940 Census about the SP and the household in which the SP lived in 1940, including state of residence. One record is included for each SP.
- The NHATS Round 5 1940 Census OP file includes information from the 1940 Census about other people living in SP's household in 1940. One record is included for each OP.
- The NHATS Round 5 1940 Census SP Geo file includes more detailed geographic information (below state). Users must provide a strong justification as to why geography is needed below state when requesting this file.

Instructions for applying to use NHATS Restricted data are available at https://www.nhats.org/researcher/data-access. To further protect the privacy of NHATS participants, some variables from the 1940 Census files have been suppressed. A list of variables, labels and brief descriptions (drawn from IPUMS) are provided in the appendix.

Complete documentation is available at the IPUMS website (Ruggles et al. 2020), including the questionnaire, https://usa.ipums.org/usa/voliii/items1940.shtml, and detailed descriptions of each of the 1940 Census variables, https://usa.ipums.org/usa. We strongly recommend users also consult the IPUMS user guide, available at https://usa.ipums.org/usa/doc.shtml.

NHATS 1940 CENSUS FILE DESCRIPTIONS

1. NHATS Round 5-1940 Census SP File

This file contains SP and household level information from the 1940 Census for NHATS SPs. The file contains 2,595 records and 73 variables.

The unique ID variable, "spid," links to all other NHATS files. Together "spid" and "pernum" uniquely identify the SP from other members of the 1940 Census household.

Household variables are located near the beginning of the file and have labels starting with "HH". Household variables include: household size, type and structure (including number of families, subfamilies, couples, mothers and fathers and an indicator of multigenerational household) and value (e.g., ownership of dwelling, monthly rent, house value and a CPI adjustment factor to 1999 dollars), and several geographic variables (region, state, urban/rural status and farm status). The Census respondent's relationship to household head is also included.

The remaining variables on this file are from the SP's person record. Some capture family-level information and others focus on the SP. Family-related variables include, information about the SP's family and subfamily (e.g. the family and subfamily to which the SP belonged, number of own family members in the household, subfamily type, relationship within subfamily), person numbers for SP's mother, father and spouse and indicators as to whether they whether the mother/father were likely step or adopted, number of own children and children under age 5 in the household, age of oldest and youngest of SP's children in the household, and number of SP's siblings in the household. In addition, SP-related variables include: relationship to household head, sex, age or age in months if <1 year, marital status, race and Hispanic ethnicity, state or country of birth, educational attainment, labor force status, industry, occupation, working hours, income, several socioeconomic status/occupation scores, whether the SP lived in the same place as five years ago, an indicator of shared surname with others in the household, and an indicator of whether the SP was the respondent for the household to the Census.

2. NHATS Round 5-1940 Census OP File

This file contains OP level information from the 1940 Census for other persons who lived with the NHATS SP in 1940. The file contains 12,476 records and 55 variables. The unique ID variable, "spid," links to all other NHATS files and to the 1940 Census SP File. The combination of "spid" and "pernum" uniquely identify each record.

OP variables on this file align with the SP-related variables on the SP file: relationship to household head, sex, age or age in months if <1 year, marital status, race and Hispanic ethnicity, state or country of birth, educational attainment, labor force status, industry, occupation, working hours, income, several socioeconomic status/occupation scores, whether the SP lived in the same place as five years ago, an indicator of shared surname with others in the household, and an indicator of whether the individual was the respondent for the household to the Census.

3. NHATS Round 5-1940 Census SP Geo File

This file contains more detailed geographic information from the 1940 Census for NHATS SPs. The file contains 2,595 records and 14 variables. The unique ID variable, "spid," links to all other NHATS files, including the 1940 Census SP File. Users interested in this file must provide additional justification as to why they need geography below state.

Geographic variables include household level information on county, metropolitan status and area, city, city population, size of place, population of place, and state economic area. In addition, details about where the SP lived 5 years ago are included on this file.

CENSUS LINKAGE AND VERIFICATION PROCEDURES

NHATS cases were linked to Public Release Version 1 of the 1940 Census using a two-step procedure. First, unique matches were identified using a machine learning algorithm. Second, cases that had more than one potential match were hand linked by trained staff. The validity and reliability of the hand linking was also assessed. The linking approach and verification procedures were developed and implemented by John Robert Warren, Fabian Pfeffer, Jonas Helgertz, and Dafeng Xu. More detail is provided below.

Available Linking Information from NHATS

For participants who were eligible for linking, we attempted to locate their record in the 1940 Census using the following information:

- Current name and Last name in 1940
- Gender, Primary race, Month and Year of birth, Age on April 1, 1940
- First and Last Name of SP's mother (if born 1924-1940 & lived with during childhood)
- First and Last Name of SP's father (if born 1924-1940 & lived with during childhood)
- First and Last Name of Other Person in SP's 1940 Household (if born 1924-1940 & lived with another adult most of childhood OR if born before 1924)
- State of Birth, State Lived in When Age 15 & State Lived in 1940

Machine Linking Algorithm

To match NHATS sample members' records to the 1940 Census we first identified potential matches. Potential matches included records that displayed identical or similar information for first and last name (in 1940 for women), age in 1940, state of birth, state of residence in 1940, and race. Deviations of ±3 years in age across data sources were allowed.

To identify the correct record from among potential matches, a machine learning algorithm was implemented to conduct probabilistic record linking. The algorithm was trained to recognize patterns of potential matches that were consistent with a true match. A modification of Feigenbaum's (2016) probit regression approach was implemented, which requires input from training data in which true matches are assumed to be known. The training data were used to calibrate the linking algorithm and to evaluate how well it performed.

To construct the training data, 500 NHATS Sample Persons were randomly selected to be linked by hand to the universe of possible 1940 matches. The universe of potential matches was limited to cases with name similarity scores (Jaro 1989; Jaro 1995; Winkler 1990; Winkler 2006) of at least 0.8, with the same gender, with year of birth within +3/-3 years, and with the same state of

birth. Unique matches were hand identified using all available information.
To calibrate the linking algorithm, a "train-test-split" procedure using the training data was implemented. In the first part of the procedure, the training data were split into two equal samples. To train the algorithm, a probit regression model was fit on one-half of the sample, and then evaluated on its out-of-sample performance on the other half. Results from the model informed the algorithm as to which, if any, of the universe of possible matches should be considered a valid link.

The algorithm declared a unique link based on (1) the probit-based predicted probability of a match and (2) the relative difference between the best and second-best possible match. By looping multiple times over a range of realistic values on both parameters, values for (1) and (2) were chosen that optimized the overall performance of the linking algorithm. Overall performance was assessed based on the algorithm's ability to minimize false positives (incorrectly linked cases) while maximizing true positives (correctly linked cases) and true negatives (correctly unlinked cases). In selecting thresholds for declaring matches, we used the Matthew's Correlation Coefficient (Chicco, 2017). After training was complete, the machine learning algorithm was applied to the full sample.

Hand Linking

For cases in which the machine linking algorithm declared one and only one high-quality match, or no high-quality matches, no hand linking was necessary.

For remaining cases, trained hand linkers were presented with a screen with NHATS information on one side and up to 10 best possible matches on the other side. Hand linkers were provided all the information provided for machine linking plus one or more parent's names or the name of another co-residing person in 1940. Hand linkers were not told which (if any) of the possible matches were declared by the machine to be a valid match and the order of the possible matches was randomized. Hand linkers were trained to identify whether there was a match and, if so, which record matched.

Verification

To assess the quality of the hand linking effort, we implemented two procedures. First, hand linkers were asked to link a random subset of cases that the machine linkage algorithm declared had one and only one high-quality match. This procedure allowed assessment of the validity of the matches—that is, how often the hand linker selected (by hand) the same match as the machine selected (via computer algorithm). There was high agreement between the hand selected and machine selected match (85% of cases); when limited to cases for which hand linkers chose a match, there was an agreement for 98% of cases. Second, we assessed the reliability of hand linking by assigning two hand linkers to the same randomly selected subset of cases. For these cases, the two independent hand linkers reached the same conclusion 90% of the time. However, most of the disagreement in these instances was due to one linker declaring a match and the other declaring no match. When both linkers declared matches, they chose the same match 99% of the time.

QUESTIONS

Questions regarding NHATS should be directed to NHATSdata@westat.com. Questions regarding access to the files should be directed to nhats-restricted-data@umich.edu. Questions about the data linkage methodology should be directed to John Robert Warren at warre046@umn.edu.

REFERENCES

Chicco, D. 2017. "Ten quick tips for machine learning in computational biology." *Biodata Mining* 10.

Jaro, M. A. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84(406):414-20.

Jaro, M. A. 1995. "Probabilistic Linkage of Large Public-Health Data Files." *Statistics in Medicine* 14(5- 7):491-98.

Feigenbaum, James J. 2016. "A Machine Learning Approach to Census Record Linking." Cambridge, MA: Department of Economics, Harvard University.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2020. " IPUMS USA: Version 10.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.

Winkler, W.E. 1990. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." American Statistical Association, 1990 Proceedings of the Section of Survey Research Methods, 354-359. http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf

Winkler, William E. 2006. "Overview of Record Linkage and Current Research Directions." Research Report Series (Statistics #2006-2), available at http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf. Washington, D.C.: U.S. Census Bureau.

Appendix. Variables in the NHATS 1940 Census Files.

Note the table below includes brief descriptions from IPUMS. Users are strongly encouraged to consult the IPUMS website for further details on how variables were constructed. https://usa.ipums.org/usa/doc.shtml. For the questionnaire, see https://usa.ipums.org/usa/voliii/items1940.shtml.

Variable	SAS Format	Label	FILE (SP/OP/GEO)	1940 Census Description / Notes
spid	8	NHATS SAMPLED PERSON ID	SP/OP/GEO	-
pernum		Person number in sample unit	SP/OP	PERNUM numbers all persons within each household consecutively in the order in which they appear on
numperhh	F_NMPRHH.	HH Number of persons in household	SP	NUMPERHH reports how many persons lived within the household.
hhtype	F_HHTYPE.	HH Household Type	SP	HHTYPE classifies all households as either family or nonfamily households. Family households are distinguished from nonfamily households using RELATE. A family household consists of a household head and one or more persons who are related to the household head by birth, marriage, or adoption and who are living together in the same household.
cpi99		HH CPI-U adjustment factor to 1999 dollars	SP	CPI99 provides the CPI-U multiplier available from the Bureau of Labor Statistics to convert dollar figures to constant 1999 dollars. This corresponds to the dollar amounts in the 2000 census, which inquired about income in 1999.
region	F_REGION.	HH Census region and division	SP	REGION identifies the region and division where the housing unit was located.
stateicp	F_STTICP.	HH State (ICPSR code)	SP	STATEICP identifies the state in which the housing unit was located, using the coding scheme developed by the Inter-University Consortium for Political and Social Research (ICPSR). The ICPSR scheme orders states first by geographic division and then alphabetically within each division.
statefip	F_STTFIP.	HH State (FIPS code)	SP	STATEFIP reports the state in which the household was located, using the Federal Information Processing Standards (FIPS) coding scheme, which orders the states alphabetically.
urban	F_URBAN.	HH Urban/rural status	SP	URBAN indicates whether a household's location was urban or rural. The term generally denotes all cities and incorporated places of 2,500+ inhabitants.
farm	F_FARM.	HH Farm status	SP	FARM identifies farm households. All group quarters are coded as non-farm, as are all housing units defined as outside the universe for FARM.
ownershp	F_OWNRSH.	HH Ownership of dwelling (tenure)	SP	OWNERSHP indicates whether the housing unit was rented or owned by its inhabitants. Housing units acquired with a mortgage or other lending arrangement(s) are classified as "owned," even if repayment was not yet completed.
rent	F_RENT.	HH Monthly contract rent	SP	RENT reports the amount of the household's monthly contract rent payment, expressed in contemporary dollars.
valueh	F_VALUEH.	HH House value	SP	VALUEH reports the value of housing units in contemporary dollars in a continuous variable.
nfams	F_NFAMS.	HH Number of families in household	SP	NFAMS is a constructed variable that counts the number of families within each unit. A "family" is any group of persons related by blood, adoption, or marriage. An unrelated individual is considered a separate family.
nsubfam	F_NSUBFM.	HH Number of subfamilies in household	SP	NSUBFAM provides the number of subfamily units within each household.
ncouples	F_NCPLES.	HH Number of couples in household	SP	NCOUPLES is a constructed variable that counts the number of married couples within each household.

nmothers	F_NMTHRS.	HH Number of mothers in household	SP	NMOTHERS is a constructed variable that counts the number of women within each household who are identified as residing with their children. Units with no mothers present are coded "0."
nfathers	F_NFTHRS.	HH Number of fathers in household	SP	NFATHERS is a constructed variable that counts the number of men within each household who are identified as residing with their children. Units with no fathers present are coded "0."
multgen	F_MLTGEN.	HH Multigenerational household	SP	MULTGEN identifies the number of distinct generations contained in each household.
respond	F_RSPOND.	HH Respondent's relationship to household head	SP	RESPOND reports how the respondent (the person who answered the enumerator's questions was related to the household head. Codes differ from RELATE.
respondt	F_RSPNDT.	Respondent indicator	SP/OP	RESPONDT identifies the person who provided most (or all) of the information about the household to the
famunit	F_FAMUNT.	Family unit membership	SP/OP	FAMUNIT indicates to which family within the housing unit each person belongs.
famsize	F_FAMSIZ.	Number of own family members in household	SP/OP	FAMSIZE counts the number of own family members residing with each individual, including the person her/himself. Persons not living with others related to them by blood, marriage/cohabitating partnership, or adoption are coded 1.
subfam	F_SUBFAM.	Subfamily membership	SP/OP	SUBFAM indicates to which subfamily (if any) within the housing unit each person belongs.
sftype	F_SFTYPE.	Subfamily type	SP/OP	SFTYPE indicates the type of subfamily (if any) to which each person belongs.
sfrelate	F_SFRELT.	Relationship within subfamily	SP/OP	SFRELATE indicates the relationship of people within their subfamily. Persons not in a subfamily are assigned a value of 0. The Census Bureau assigns a "reference person" to each subfamily. In married-couple subfamilies, this is the husband; in parent-child subfamilies, this is the parent.
momloc	F_MOMLOC.	Mother's location in the household	SP/OP	MOMLOC is a constructed variable that indicates whether the person's mother lived in the same household and, if so, gives the person number of the mother (PERNUM). The method by which probable child-mother links are identified is described in MOMRULE_HIST.
stepmom	F_STPMOM.	Probable step/adopted mother	SP/OP	STEPMOM reports whether a person's mother, as identified by MOMLOC, was likely to have been the person's stepmother or adoptive mother.
momrule_hist	F_MMRLHS.	Rule for linking mother	SP/OP	MOMRULE_HIST reports why MOMLOC linked the person to a probable mother.
poploc	F_POPLOC.	Father's location in the household	SP/OP	POPLOC is a constructed variable that indicates whether the person's father lived in the same household and, if so, gives the person number of the father (PERNUM). The method by which probable child-father links are identified is described in POPRULE_HIST.
steppop	F_STPPOP.	Probable step/adopted father	SP/OP	STEPPOP reports whether a person's father, as identified by POPLOC, was likely to have been the person's stepfather or adoptive father.
poprule_hist	F_PPRLHS.	Rule for linking father	SP/OP	POPRULE_HIST reports why POPLOC linked the person to a probable father.
sploc	F_SPLOC.	Spouse's location in household	SP/OP	SPLOC is a constructed variable that indicates whether the person's spouse lived in the same household and, if so, gives the person number (PERNUM) of the spouse.
sprule_hist	F_SPRLHS.	Rule for linking spouse	SP/OP	SPRULE_HIST reports why the IPUMS variable SPLOC linked the person to a probable spouse.
nchild	F_NCHILD.	Number of own children in the household	SP/OP	NCHILD counts the number of own children (of any age or marital status) residing with each individual. NCHILD includes step-children and adopted children as well as biological children. Persons with no children present are coded "0."
nchlt5	NCHLT5F.	Number of own children under age 5 in household	SP/OP	NCHLT5 counts the number of own children age 4 and under residing with each individual. NCHLT5 includes step-children and adopted children as well as biological children. Persons with no children under 5 present are coded "0."
nsibs	F_NSIBS.	Number of own siblings in household	SP/OP	NSIBS counts the number of own siblings (including half-siblings, step-siblings, and adopted siblings) residing with each individual. Persons with no siblings present are coded "0."

eldch	F_ELDCH.	Age of eldest own child in household	SP/OP	ELDCH reports the age of the eldest own child (if any) residing with each individual, regardless of the child's age or marital status. ELDCH includes step-children and adopted children as well as biological children. The highest legitimate age for ELDCH is 98. Persons with no children present are coded 99.
yngch	F_YNGCH.	Age of youngest own child in household	SP/OP	YNGCH reports the age of the youngest own child (if any) residing with each individual, regardless of the child's age or marital status. YNGCH includes step-children and adopted children as well as biological children. The highest legitimate age for YNGCH is 98. Persons with no children present are coded 99.
relate	F_RELATE.	Relationship to household head	SP/OP	RELATE describes an individual's relationship to the head of household or householder. The relationship codes are divided into two categories: relatives (codes 1-10) and non-relatives (codes 11-13).
sex	F_SEX.	Sex	SP/OP	SEX reports whether the person was male or female.
age	F_AGE.	Age	SP/OP	AGE reports the person's age in years as of the last birthday.
agemonth	F_AGEMTH.	Age in months	SP/OP	AGEMONTH reports the age, in months, of persons less than one year old (AGE = 0) on census day.
marst	F_MARST.	Marital status	SP/OP	MARST gives each person's current marital status.
race	F_RACE.	Race	SP/OP	The census enumerator was responsible for categorizing persons by RACE and was not specifically instructed to ask the individual his or her race.
hispan	F_HISPAN.	Hispanic origin	SP/OP	HISPAN identifies persons of Hispanic/Spanish/Latino origin and classifies them according to their country of origin when possible. The variable was inferred from other variables. See HISPRULE.
bpl	F_BPL.	Birthplace	SP/OP	BPL indicates the U.S. state, the outlying U.S. area or territory, or the foreign country where the person was born.
spanname	F_SPNNME.	Spanish surname	SP/OP	SPANNAME identifies persons with Spanish surnames, based on comparisons of surnames with lists of Spanish surnames.
hisprule	F_HSPRUL.	Hispanic origin rule	SP/OP	HISPRULE reports why a person was coded as Spanish/Hispanic/Latino.
school	F_SCHOOL.	School attendance	SP/OP	SCHOOL indicates whether the respondent attended school during the month prior to the Census (March 1-April 1).
higrade	F_HIGRAD.	Highest grade of schooling	SP/OP	HIGRADE reports the highest grade of school completed by the respondent.
educ	F_EDUC.	Educational attainment	SP/OP	EDUC indicates respondents' educational attainment, as measured by the highest year of school or degree
empstat	F_EMPSTT.	Employment status	SP/OP	EMPSTAT indicates whether the respondent was a part of the labor force -- working or seeking work -- and, if so, whether the person was currently unemployed.
labforce	F_LABFRC.	Labor force status	SP/OP	LABFORCE is a dichotomous variable indicating whether a person participated in the labor force.
occ		Occupation	SP/OP	OCC reports the person's primary occupation, coded into a contemporary census classification scheme. For details see: https://usa.ipums.org/usa/volii/occ1940.shtml
occ1950	OCC1950F.	Occupation, 1950 basis	SP/OP	OCC1950 applies the 1950 Census Bureau occupational classification system to occupational data, to enhance comparability across years.
ind		Industry	SP/OP	IND is an un-recoded variable that reports the type of industry in which the person performed an occupation, which is recorded in the variables OCC (Occupation) and OCC1950 (Occupation, 1950 basis). In census usage, "industry" currently refers to work setting and economic sector, as opposed to the worker's specific technical function, or "occupation".
ind1950	IND1950F.	Industry, 1950 basis	SP/OP	IND1950 recodes information about industry into the 1950 Census Bureau industrial classification system.
classwkr	F_CLSWKR.	Class of worker	SP/OP	CLASSWKR indicates whether respondents worked for their own enterprise(s) or for someone else as employees. Workers with multiple sources of employment were classified according to the work relationship in which they spent the most time during the reference day or week.
wkswork1	WKSWRK1F.	Weeks worked last year	SP/OP	WKSWORK1 reports the number of weeks that the respondent worked for profit, pay, or as an unpaid family worker during the previous calendar year

wkswork2	WKSWRK2F.	Weeks worked last year, intervalled	SP/OP	WKSWORK2 reports in intervals the number of weeks that the respondent worked for profit, pay, or as an unpaid family worker during the previous calendar year.
hrswork1	HRSWRK1F.	Hours worked last week	SP/OP	HRSWORK1 reports the total number of hours the respondent was at work during the previous week. For employers and the self-employed, this includes all hours spent attending to their operation(s) or enterprise(s). For employees, it is the number of hours they spent at work. For unpaid family workers, it is the number of hours they spent doing work directly related to the family business or farm.
hrswork2	HRSWRK2F.	Hours worked last week, intervalled	SP/OP	HRSWORK2 reports in intervals the total number of hours the respondent was at work during the previous week.
durunemp	F_DRUNEM.	Continuous weeks unemployed	SP/OP	DURUNEMP reports how many consecutive weeks had elapsed since each currently-unemployed respondent was last employed (i.e., how many weeks had the person been without a job and looking for one).
incwage	F_INCWAG.	Wage and salary income	SP/OP	INCWAGE reports each respondent's total pre-tax wage and salary income - that is, money received as an employee - for the previous year. Amounts are expressed in contemporary dollars.
occscore		Occupational income score	SP/OP	OCCSCORE is a constructed variable that assigns occupational income scores to each occupation based on OCC1950.
sei	F_SEI.	Duncan Socioeconomic Index	SP/OP	SEI is a constructed measure that assigns a Duncan Socioeconomic Index (SEI) score to each occupation using the 1950 occupational classification scheme available in the OCC1950 variable. The SEI is a measure of occupational status based upon the income level and educational attainment associated with each occupation in 1950. See https://usa.ipums.org/usa/chapter4/sei_note.shtml
presgl	F_PRESGL.	Occupational prestige score, Siegel	SP/OP	PRESGL is a constructed variable that assigns a Siegel prestige score to each occupation using the occupational classification scheme available in OCC1950 variable. See https://usa.ipums.org/usa/chapter4/sei_note.shtml
erscor50	ERSCR50F.	Occupational earnings score, 1950 basis	SP/OP	ERSCOR50 is a constructed variable that assigns a measure of the median earned income for each occupation using the 1950 occupational classification scheme available in the OCC1950 variable.
edscor50	EDSCR50F.	Occupational education score, 1950 basis	SP/OP	EDSCOR50 is a constructed variable indicating the percentage of people in the respondent's occupational category who had completed one or more years of college.
npboss50	NPBOS50F.	Nam-Powers-Boyd occupational status score, 1950 basis	SP/OP	NPBOSS50 is a constructed 4-digit numeric variable that assigns a Nam-Powers-Boyd occupational status score to each occupation using the occupational classification scheme available in OCC1950 variable. https://usa.ipums.org/usa/chapter4/sei_note.shtml
migrate5	MIGRAT5F.	Migration status, 5 years	SP/OP	MIGRATE5 indicates whether a person age 5+ had changed residence since a reference point 5 years ago. Specifically, individuals were asked if they had lived in the "same house" (non-movers) or a "different house" (movers) five years earlier.
sursim	F_SURSIM.	Surname similarity	SP/OP	SURSIM assigns the same code to all persons within each household who had the same surname.
countyicp		HH County (ICPSR code)	GEO	COUNTYICP identifies the county where the household was enumerated, using the Inter-University Consortium for Political and Social Research (ICPSR) coding scheme. ICPSR county codes are generally ordered alphabetically by county name within states. COUNTYFIP codes are state-dependent; they must be combined with state codes to distinguish counties located in different states. For details see: https://usa.ipums.org/usa/volii/ICPSR.shtml
metro	F_METRO.	HH Metropolitan status	GEO	METRO indicates whether the household resided within a metropolitan area and, for households in metropolitan areas, whether the household resided within or outside of a central/principal city. A metropolitan area, or metro area, is a region consisting of a large urban core together with surrounding communities that have a high degree of economic and social integration with the urban core.
metarea	FMETAREA.	HH Metropolitan area	GEO	For residents of metro areas, METAREA identifies the metro area of residence,
city	F_CITY.	HH City	GEO	CITY identifies the city of residence for households located in identifiable cities. The cities identified by CITY are generally consistent with U.S. Census "place" definitions.

citypop	F_CTYPOP.	HH City population	GEO	CITYPOP reports the population, in hundreds, for all identifiable cities collected in the 1940 Census.
sizepl	F_SIZEPL.	HH Size of place	GEO	SIZEPL is a recode of CITYPOP, grouping places of similar sizes.
urbpop		HH Population of urban places	GEO	URBPOP gives the population, in hundreds, of places considered "urban" according to the Census Bureau's 1930-1940 definition.
sea	F_SEA.	HH State Economic Area	GEO	SEA identifies the State Economic Area in which the household is located. State Economic Areas are single counties or groups of contiguous counties within the same state that had similar economic characteristics in 1950. The concept was retrospectively applied to 1940.
cntry	F_CNTRY.	HH Country	GEO	CNTRY gives the country from which the sample was drawn. The codes assigned to each country are those used by the UN Statistics Division and the ISO (International Organization for Standardization).
migplac5	MGPLAC5F.	State or country of residence 5 years ago	GEO	MIGPLAC5 identifies the U.S. state, outlying territory, or foreign country where the individual who changed residences lived 5 years ago.
migtype5	MGTYPE5F.	Metropolitan status 5 years ago	GEO	MIGTYPE5 indicates whether the individual who changed residences lived in a metropolitan area five years ago and in the largest city/cities ("the central cities") in the metropolitan area.
sameplac	F_SAMPLC.	Lived same incorporated place 5 years ago	GEO	SAMEPLAC indicates whether individuals were living in the same community ("place") on April 1, 1935 as they were living at the time of enumeration.
samesea5	SAMSEA5F.	Lived same SEA 5 years ago	GEO	SAMESEA5 reports whether the respondent lived in the same State Economic Area (SEA) on April 1, 1935 as the person was living in at the time of the 1940 enumeration.