Documentation for the 2003-2015 PSID Event History Calendar (EHC) Between-Wave Moves File

July 2018

https://psidonline.isr.umich.edu
psidhelp@umich.edu
734-763-5166

Panel Study of Income Dynamics
Economic Behavior Program
Survey Research Center
Institute for Social Research
University of Michigan
PO Box 1248
Ann Arbor, MI 48106-1248

Table of Contents

Introduction: 2003-2015 PSID EHC Between-Wave Moves Data

The 2003-2015 PSID EHC Between-Wave Moves file currently contains PSID addresses from the Event History Calendar, for waves 2003-2015. EHC address data were geocoded using the SAS proc geocode process. Census geocodes on this file are from the 2010 Census. The FIPS county identifiers are current as of the time of release.

Difference between "PSID Geocode Match" and "Between-Wave Moves" Files

Unlike the Geocode Match file, where there is only one record per family unit, the EHC Between-Wave Moves file has up to six records for each family. Additionally, while the Geocode Match file includes a single address for the family unit, the records in the EHC Move file are for the Head. That is, no matter who the Respondent is, the question about current and past residences are asked about the Head only.

Finally, there are instances when a family unit's current address data on the EHC Between-Wave Moves file may not match the address data in the Geocode Match file. This happens for a myriad of reasons. For one, the Geocode Match File data are derived from the PSID Address file which is used for PSID Respondent payments and mailings while EHC Between-Wave Moves data come from responses to the current and past residence question. As a result, a permanent address, where a Respondent wants payment or mailings sent may differ from the physical address where Respondent actually resides. We also have Respondents who want payments and mailings to go to a parent or other relative's address. Additionally, Respondents may provide a P.O. Box for payments but provide an actual physical address for the EHC.

Address Data Used for the Between-Wave Moves File

The geocoded addresses that are included in this file are from the PSID Event History Calendar (EHC) files, confidential sets of addresses kept separately from the main PSID data, for the 2003-2015 waves. Please refer to Table 1, below, which lists the number of addresses and number of families per wave.

At the beginning of Section BC: Employment, Respondents are asked about current and previous residences in the past two years for the Head, with a maximum of six records per wave (please refer to Item 1 below). The question is asked, as follows:

1. [*PRELOADED:* IWER: Residence timeline #1 has been preloaded with [HEAD]'s [P2YEAR] residence. CLICK on address label (left-column) and PROBE whether this is still [his / her] current residence. VERIFY and EDIT Start Date and Address as needed. / *ALL OTHERS*: What is the street address and move-in date of [your / [HEAD]'s] current residence?]
2. [*PRELOADED*: If needed:] [Have you / Has [he / she]] lived anywhere else since January [P2YEAR]? IF YES: Please tell me the address of each of those residences and the dates [you / he / she] lived there.

Item 1: *Example screen shot from the 2017 Blaise instrument using testing data*

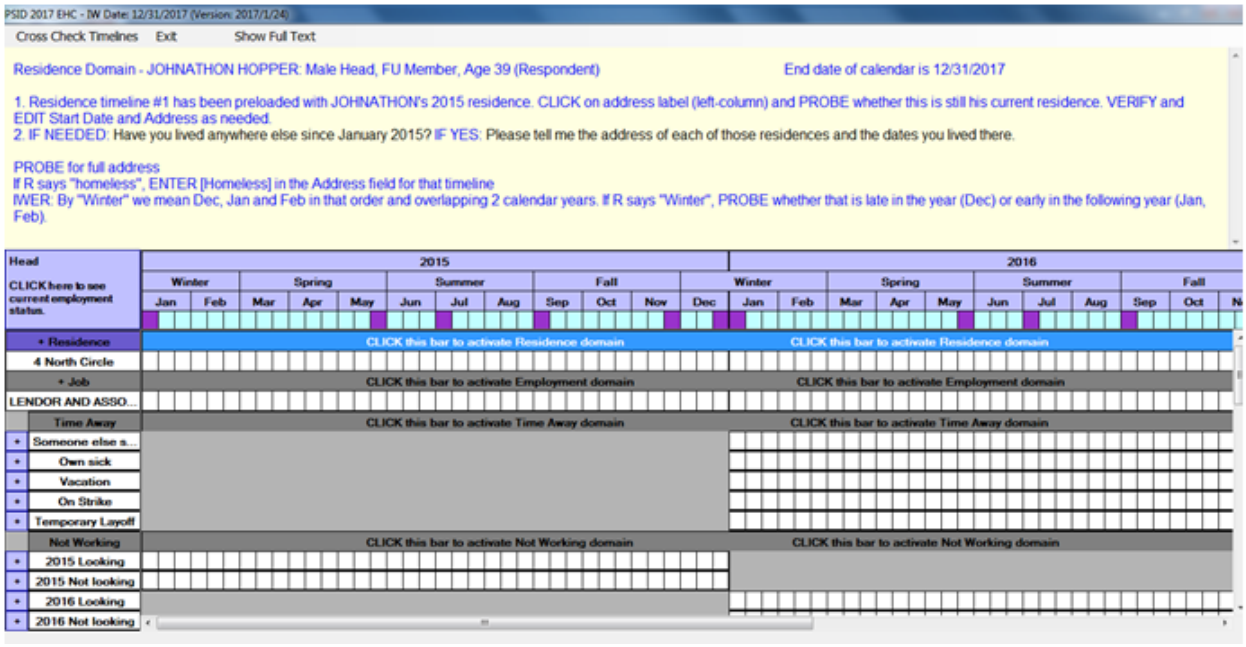

Table 1: Number of Families and EHC Addresses Per Wave

Interview year	Number of addresses	Number of Families
2003	11,786	7,822
2005	12,399	8,002
2007	13,214	8,289
2009	13,410	8,690
2011	14,418	8,907
2013	14,536	9,063
2015	14,422	9,048

Data Characteristics of the Between-Wave Moves File

As noted above, Respondents were asked to either confirm or provide current residence as well as any previous residences since January, two years prior, for the Head. Up to eleven "moves"/residences were recorded.

The EHC Between-Wave Moves data, by the nature of the question asked, are complex. We are relying on Respondents to recall previous addresses for herself/himself, if she/he is the Head, or, more challenging yet, for someone else if she/he is not the Head. As a result, the data file has some level of incomplete or inaccurate data, even with data cleaning efforts. Furthermore, there can be timeline inconsistencies whereby Respondents may not remember the correct dates and order of moves or miss a move altogether.

Address Cleaning Process for the Between-Wave Moves File

The input to the geocode process is string variable (called EVENT in the source data files), which is street number, apartment, city, state, zip, and country -- a comma delimited string. The input string needs to be parsed into 5 variables (street, city, state, zip, country) before geocoding the address.

Due to missing data and data reporting errors, the quality of results would not be acceptable if input string is simply split by the separator (comma) without extensive data cleaning. The processes is, to identify how many parts it should be split into and which parts go to which variables, is very challenging. Below are data cleaning and standardizing steps taken before the geocoding process.

The data cleaning processes started with removing or replacing some non-alphabetic characters with space. That includes leading # or @, leading and trailing blanks, and consecutive blanks, and) (` . : [ ] ;. Then extra commas and spaces were removed. Then text ‘,USA’ was added to ending digits (presumptive ZIP code).

Identifying foreign country address is an important task. Only USA addresses are geocoded. Frist, we flagged foreign address by searching for a list of country names. Then the ending ‘,USA’ in a foreign address is removed. The foreign address records would not be further parsed or geocoded.

Another big challenge was to identify apartment number in address string. Apartment number was not separated from street address by comma in the input data. Geocoding can only match the street address without apartment number, otherwise it would return non-matches or inaccurate results. Keywords like ‘APT’, ‘UNIT’, ‘BLDG’, ‘PLAZA’, ‘SUITE’, ’#’, ‘FLOOR’, ‘ROOM’, ‘LOT’, ‘SPACE’, ‘SP’, ‘RM’, ‘DUPLEX’, ‘HOUSE’, ‘REAR’, ’BASEMENT’ were used to locate and insert comma before apartment number.

Other cleaning steps include replacing invalid zip values with ‘00000’, removing ‘DK’, replacing ‘N E’ with ‘NE’, etc. Some manual fixes were also preformed if irregular cases were discovered along the development of the data cleaning processes.

The cleaned input variable (EVENT) had the standardized form: street, [aptnum,] city, state, zip [,USA] and was ready to be parsed into the correct fields if it was a USA address and contains 4 or 5 field separators (commas).

After data parsing, variable STATE was corrected with SASHELP.ZIPCODE to fix invalid data.

The Geocoding Process for the Between-Wave Moves File

The geocoding process for this file follows the same process used in the 2010 PSID Geocode Match file. That is, EHC data were reviewed and cleaned, followed by use of the SAS 9.4 proc geocode process. We imported the latest TIGER/Line shape files for all states from the Census Bureau. These files can be found here: https://www.census.gov/geo/maps-data/data/tiger-line.html.

When an exact match is not found, SAS will use other values to get the closest match. The accuracy of that match is coded EHCV23. Another accuracy variable, EHCV21, "SAS Numeric Quality of Match", gives a numeric representation of how good the geocode match is. The score is calculated based on EHCV22, "SAS Match Tokens." Each token within EHCV22 has a numeric value that is used to calculate EHCV21. For example, if EHCV22 contains "AD ZC NM" then that means the street name, zip code, and house number matched. The values in EHCV21 equal the sum of 20 for AD, 15 for ZC and 10 for NM for a total score of 45.