Understanding Society: Connecting family members and households over time using the Family Matrix (xhhrel)

USER GUIDE

Version 1, September 2022

Contents

1. Introduction

1.1 What is the family matrix (xhhrel)?

Understanding Society provides an unprecedented window on the interrelations between different family members within and across households and generations. To learn more about the Study, take a look at the study overview and the user guide.

Making family connections across different households over time can be difficult, so to help researchers link families we have created a family matrix data file **xhhrel**. This data file identifies all family relationships across the Study, allowing researchers to see where people and households connect. As this is a household longitudinal survey, information is collected about its core sample members as well as everybody they are currently living with in the same household, every year from the time they were selected into the Study. Each year information is collected about the relationships of every household member and how they relate to each other. Attaching information about these family members to that of the respondents enables analysts to expand the type of research questions they are able to answer. For example, do individuals tend to partner with those with similar educational qualifications? What is the degree of social mobility in the UK? To facilitate research on such family connections and help data users link families more easily the family matrix (**xhhrel**) identifies all family relationships consistently across the Study.

The **xhhrel** file creates an individual level cross-wave file of all sample members (those who were ever enumerated as part of the Study) that contain familial relationship identifiers reported over the survey period for each sample member. This file also contains an origin household identifier variable (**osm_hh**) which identifies the household they come from, so that sample members who are connected with each other can be identified (either because they were co-resident at some point or were co-resident with individuals who were co-resident with each other).

1.2 Why should you use the family matrix (xhhrel)?

The information about the relationship between any two household members is provided in separate wave specific files, **w_egoalt,** where w refers to the wave (a for Wave 1, b for Wave 2 and so on). Each row in this file includes the **pidp** (the unique cross-wave personal identifier) of a household member, as well as the **pidp** of every other household member and their relationship to each other. Using this information, the data user can produce the **pidp** of the parent, child, sibling etc. Users should note that **pidp** of some of the key co-resident relationships, parents, grandparents, spouses, and partners, are already provided in wave specific files, **w_indall**. But this information is only available for co-resident household members. So for example, if a child was living with their parents in a specific wave, the information about the identity of their parents will be available in that wave specific file. If the child moves out of their parents' household in a subsequent wave, information about their parents' identifiers will not be available in that wave specific file. Therefore, to identify parents, irrespective of the wave they were co-resident in, data users would have to combine information across multiple wave specific data files to access this information. In other words, while this information is available it is not available in an easy-to-use format. The new **xhhrel** file makes this process much easier by identifying all family relations across all waves.

Secondly, information about a few key (co-resident) relationships (such as, parents, grandparents, spouses, and partners) are already provided in these wave specific files (as stated above), so the identifiers of other relationships, such as siblings, will have to be coded by the data user if they decide not to use the **xhhrel** file.

Finally, in a few cases, the relationship between a pair of sample members is coded differently in multiple waves (for example, father in Wave 1 and brother in Wave 2) either due to reporting or coding error. In such cases, if the **xhhrel** file is not used, information about the relationship of these two sample members will need to be considered across all waves and decisions taken about their "correct" relationship.

2. Data structure

2.1 Contents of the family matrix xhhrel file

This file compiles existing Understanding Society main survey data, particularly from the **egoalt** and **indall** files, and includes every sample member ever enumerated as part of the Study. It is in wide format and unique on **pidp**, i.e., each row represents a separate sample member. The variables included are the **pidp** of the sample member, the **osm_hh** they come from and the **pidps** of other sample members who are connected via different relationships. The relationships listed are static so there are no start/end dates. For example, if a person was married twice and both spouses were enumerated as part of the study, then the **pidps** of both partners will be provided without any information about the duration of these relationships. The file, **xhhrel**, contains identifiers of partners, parents/children (biological, adoptive, step, foster, grand, in-law), siblings (biological, half, adoptive, step, foster, in-law), aunts/uncles, niece/nephews and cousins. These relationship variables follow this naming convention: the root variable name is prefixed with the following prefixes each signifying the relationship.

Table 1: Naming convention of relationship identifier variables

prefix	relationship	prefix	relationship
`pt`	partner	`gp`	grandparent
`bp`	biological parent	`gc`	grandchild
`bc`	biological child	`bsb`	biological sibling
`ap`	adoptive parent	`hsb`	half sibling
`ac`	adopted child	`asb`	adopted sibling
`sp`	stepparent	`ssb`	stepsibling
`sc`	stepchild	`fsb`	foster sibling
`fp`	foster parent	`lsb`	sibling-in-law
`fc`	foster child	`au`	aunt/uncle
`lp`	parent-in-law	`nn`	niece/nephew
`lc`	child-in-law	`cn`	cousin

In addition to the **osm_hh**, **pidp** and the relationship identifiers, the data file, **xhhrel**, includes some key sample member information variables, such as sex, year of birth and sample origin, and some summary relationship variables (see Table 2 for a complete list).

Table 2: List of variables in xhhrel

Variables	Description
`osm_hh`	Original Sample Member (OSM) household identifier. **osm_hh** is generated from the data collected in the first wave a household is enumerated in and is used to identify sample members who are connected with each other (either because they were co-resident at some point or were co-resident with individuals who were co-resident with each other).
`pidp`	Cross-wave unique person identifier.
`osm_hh_size`	Total number of people in **osm_hh**.
`osm_hh_head`	Flag identifying the head of the **osm_hh**.
`sex, doby_dv, hhorig, sampst, dcsedfl_dv dcsedw_dv, lwenum_dv_bh, lwenum_dv, lwintvd_dv_bh, lwintvd_dv`	Sampling and basic demographic information merged from cross-wave files.
r`x_N`	Total number of specified relationships identified in the survey, where *r* is the relationship prefix.
r`x_pidp_`*n*	Cross-wave unique person identifier of the specified relation number *n,* where *r* is the relationship prefix.
r`x_sex_`*n*	Sex of the specified relation number *n*, where *r* is the relationship prefix.
`ptx_cr_`*n*	Flag identifying whether partner number *n* is the current partner based on the latest wave of data (i.e., Wave 11).
`gpx_pf_`*n*	Flag identifying which parent the grandparent number *n* is related to (if known).

2.2 Missing values and data inconsistencies

Missing values are flagged in the same way as the main study, using only -8 (inapplicable) and -9 (missing).

Data inconsistencies that could not be clearly resolved at this time have been highlighted (ranging from <0.01% to 3.17% of the sample). We will investigate these inconsistencies as we receive new data and try to resolve as many as possible. In addition, to the variables listed in Table 2, this data file also contains variables that flag the different types of data inconsistencies identified (see Table 3).

Table 3: list of data inconsistency flag variables

Variables	Description
`apx_rr_ef`	1 or more adoptive parent/adopted child inverse relationships do not match
`bpx_rr_ef`	1 or more biological parent/biological child inverse relationships do not match
`bsbx_rr_ef`	1 or more biological sibling inverse relationships do not match
`fpx_rr_ef`	1 or more foster parent/foster child inverse relationships do not match
`gpx_rr_ef`	1 or more grandparent/grandchild inverse relationships do not match
`hsbx_rr_ef`	1 or more half sibling inverse relationships do not match
`lpx_rr_ef`	1 or more parent-in-law/child-in-law inverse relationships do not match
`lsbx_rr_ef`	1 or more sibling-in-law inverse relationships do not match
`spx_rr_ef`	1 or more stepparent/stepchild inverse relationships do not match
`bpx_ef`	more than 2 biological parent(s) listed
`mr_ef`	1 or more relationships reported for the same person. Some of these can be legitimate (e.g., reporting one person as both adoptive and stepparents).

2.3 Linking to the main survey

The file, **xhhrel**, only includes information about the relationships and some sample member demographic information, such as sex, year of birth, etc. To get additional information about the sample members as well as that about their relatives, this data file will need to be combined with Understanding Society’s main survey data files where that information is available. As this file includes **pidp**, this variable can be used to link with all main survey individual level files (which also include **pidp**). However, to make it easier to do this, we have provided syntax for some common procedures we think users might want to do with the file. See Section 4.2.

3. Data access, citation, help and support

This data file is provided as part of Understanding Society main survey data release and this document is part of the main survey user guide. Visit these webpages to learn how to access the data and cite the data and user guide.

In this short video, Nicole James explains what the family matrix is, how it works and how it can help researchers who want to identify families across households and across time.

If users have further questions about the **xhhrel** file, please contact User Support via email usersupport@understandingsociety.ac.uk or through our User Forum.

4. Stata syntax for some common procedures

4.1 Identifying individuals who have specific relatives in the study

- Identify individuals with parents in the study.

  ```
  browse if px_N>0
  ```

- Identify individuals with grandparents in the study.

  ```
  browse if gpx_N>0
  ```

- Identify individuals with parents and grandparents in the study.

  ```
  browse if px_N>0 & gpx_N>0
  ```

4.2 Using xhhrel to match information of sample member to that of specific relatives

The purpose of this syntax is to attach information of specific family members using the main survey data and the family identifiers from the family matrix.

```
/* Step 1: Write the complete file-path of the location where the downloaded data is stored
within the double quotes in the line below. */
global curr_release ""

/* Step 2: Open the family matrix data file, choose the family relationship whose information
you need to match and save it as "temp.dta"
In this example we are using bpx_pidp_1 which is the pidp of the first biological parent. */
use "${curr_release}\ukhls\xhhrel", clear
keep if bpx_pidp_1>0 & !mi(bpx_pidp_1)
save "temp", replace

/* For UKHLS data files (Waves 1-11) */
forvalues i = 1/11 {
        local w = substr("abcdefghijklmnopqrstuvwxyz",`i',1) + "_"

        /* Step 3: Open the data file from which the information is to be matched
        and choose the variables you want to keep. In this example, we are using jbsemp.
        Note: This syntax uses user-generated command "isvar", which can be
        installed using the ssc install command. */

        use "$curr_release/`w'indresp", clear
        isvar `w'jbsemp // specify variables
        keep pidp `r(varlist)'

        /* Step 4: Rename pidp to that identifier */
        rename pidp bpx_pidp_1

        /* Step 5: Rename the variables with a suffix to show these refer to that
        family member */
        rename `w'* `w'*_bpx_1

        /* Step 6: Merge with the family matrix */
        merge 1:m bpx_pidp_1 using "temp", keep(2 3) nogen
        isid pidp
        order `w'*_bpx_1, last

        save "temp", replace
}
```