Data sourcing, processing and preparation

See <http://doi.org/10.5255/UKDA-SN-853082> for processing and checks on the original Wrigley et al dataset.

See also George Alter, Gill Newton, and Jim Oeppen, "Re-introducing the Cambridge Group Family Reconstitutions," Historical Life Course Studies (in review).

The following describes how UKDA-SN-853082 was transformed for this data collection.

**A: Converting the Cambridge Group Reconstitutions into IDS**

Transferring data into IDS involves thinking about data in a different way than working with data in a statistical package, like SPSS or Stata. IDS uses an “entity-attribute-value” model to describe data. In IDS every piece of information or “datum” is a separate record. A record always includes an identifier pointing to an “entity,” which is either a person (ID\_I) or a “context” (ID\_C). Every record includes a Type or Relation, designating the kind of information it provides, which may be the date of an event, the Value of an “attribute,” or the relationship between two people.

The main steps in transferring the reconstitutions from the relational database into IDS were:

1. Add IDs for individuals and contexts

The nine data tables in the Reconstitutions were linked to each other by a single identifier, parfrf, but IDS requires an identifier for every individual.

We created individual IDs by adding a role suffix to the parfrf.

• Wife: parfrf + “W”

• Husband: parfrf + “H”

• Child: parfrf + “C” + “xx” where xx is ChildNumber

• Wife’s father: parfrf + “WF”

• Wife’s mother: parfrf + “WM”

• Husband’s father: parfrf + “HF”

• Husband’s mother: parfrf + “HM”

Columns were added to the nine Cambridge Group data tables so that IDs for all of the individuals mentioned in the table were included. The parfrf was also used as a “context ID” for the marital union. Context is a flexible concept that can be used in IDS for any way of grouping people together. A context can be a physical location, such as a house or apartment, or a social grouping, like a household. A context can also be an artifact of the way in which information is arranged in a source document, like a page in a census or population register. An FRF in family reconstitution corresponds to a conjugal unit consisting of a husband, wife, and children, and identifying each FRF as a “Union” in IDS is convenient for data extraction software.

1. Births/baptisms and deaths/burials

Most of the original reconstitutions provided dates for baptisms and burials, some gave dates for birth and death, and a small number included both dates. Baptisms were usually a few days after birth, but the delay is sometimes much longer. In a few families, four or five children were baptized on the same day, which probably means that siblings of different ages were baptized together. IDS includes separate types for birth (BIRTH\_DATE), baptism (BAPTISM\_DATE), death (DEATH\_DATE) and burial (FUNERAL\_DATE), and all of the dates in the FRFs were included in the IDS file.

1. Convert “weightings” to Estimations

Information about the quality of dates in the family reconstitutions was coded into numeric codes called “weightings” in the Cambridge Group (Wrigley et al., 1997, pp. 563-568). For example, if the original document provided the month and year but not the day of a death, the method for computing the day is given in the “death weighting” variable. There are 120 weighting codes in the Cambridge Group family reconstitution database, and the meanings of some codes have been lost. These codes were ordered so that higher numbers indicated less reliability. Weightings were mapped to the IDS Estimation field, which describes the precision of dates. If no weighting is associated with a date, we assume the source provided day, month, and year and designate it as “Exact” in IDS. If only the year was given, we assign an Estimation of “Estimated (dd/mm).” Some weightings were translated to “Before this date” or “After this date.” For example, sometimes a woman remarried, but the date of death of her first husband is not known. In these cases the date of death of the first husband may be listed as the date of her second marriage with Estimation as “Before this date.” (See also Wrigley et al. (1997, p. 576) for dates father’s death inferred from posthumous births.)

1. Convert “flags” to Types

“Flags” were used for various types of less common information. There are flags for widowhood, religious denominations, illegitimacy, and paupers. Flags were translated into IDS information Types. For example, senior/junior and elder/younger were recorded in IDS Type “SUFFIX\_NAME.” Columns were added to the nine Cambridge Group data tables for Pauper, Posthumous, Name suffix, and Religion, because these Types could not be recorded in existing columns. We did not transcribe flags when they could be inferred from information available elsewhere in the database, such as marital status at death.

1. Map columns to IDS Types and IDS Transposer

The IDS Transposer is a web service that converts rectangular data files into IDS format files (Klancher Merchant & Alter, 2017). The IDS Transposer requires two “mapping” files describing how the original data files will be represented in IDS. We give a brief overview here and refer readers to Klancher Merchant and Alter (2017) for advice on the construction of the mapping files. The Entity mapping file associates columns in the input files with attributes and events in the IDS INDIVIDUAL and CONTEXT tables. Each IDS record must have a Type identifying the kind of information it contains and an ID pointing to an individual or a context. INDIVIDUAL and CONTEXT records may have a Value (such as “male” or “female” for Type=Sex), and most records have a timestamp showing when an event occurred or the attribute was observed. The Relationship mapping file controls the creation of records for the INDIV\_INDIV, INDIV\_CONTEXT, and CONTEXT\_CONTEXT tables. These tables describe a Relation between two entities, i.e. individuals or contexts. Relationships in the INDIV\_INDIV table are always given in both directions, e.g. “mother of” and “child of” (Alter & Mandemakers, 2014). As mentioned above, we consider each FRF a marital union. A record was created for each FRF in the CONTEXT table with Type=”Union” and timestamp set to the date of marriage. Each person in the FRF was linked to the Union in the INDIV\_CONTEXT table with a relation of “Wife,” “Husband,” or “Child.” Most of the work of transferring data from the nine tables in the Cambridge Group reconstitutions database to the five IDS tables was done by the IDS Transposer, but several additional processing steps were needed.

1. Harmonize individual IDs

Wherever possible, individuals appearing in more than one FRF were assigned the same ID everywhere in the IDS database. References linking individuals across FRFs are available in the original database. The Children table includes a column showing the FRF of the first marriage of that child, and the Husbands and Wives tables have columns for previous and next marriages. These references were used to create an ID crosswalk, and individuals were assigned the ID that occurred earliest in time in the database.

1. Add parish-level context information

We use the CONTEXT table to add information that is common to every FRF in a parish, such as county, “quality” years, and Sample Groups, which are discussed below. Since IDS uses a relational model, “quality” years and Sample Groups are recorded once in the CONTEXT table and linked to each FRF through a “Union in parish” relationship in the CONTEXT\_CONTEXT table.

1. Other processing

A number of other minor changes were made to the tables produced by the IDS Transposer. In some situations, the IDS Transposer created empty records that were deleted. We also corrected some errors in the database. We ran a suite of error checks on the IDS database to detect inconsistencies, such as death dates that precede birth or marriage dates or births to mothers younger than 15 or older than 50. A list of these known problems is included with the IDS database.

**B: From IDS to Episodes**

Data in IDS format must be converted to a rectangular data array for analysis by standard statistical packages. Longitudinal data described in IDS needs to be transformed into episodes – segments of time in a fertility history. Each episode has start and end dates. Episodes that do not end in a birth are considered “right censored” in event history analysis. Episodes are divided (left and right truncated) when a variable changes, so that all explanatory variables are constant within an episode. We use a two-step process developed by Luciana Quaranta to move from IDS to an episodes file (Quaranta, 2015, 2016). The first step is creating a “chronicle” file from the IDS. A “chronicle” file uses an “entity-attribute-value” approach very similar in structure to IDS, but the attributes may be complex time-varying measures, such as “number of co-resident older siblings.” Attributes in the Chronicle file will become variables in the Episode file used for analysis. The Chronicle file was created using SQL in Microsoft Access. The Chronicle file was imported to Stata where it was converted to episodes by a modified version of Quaranta’s Episode File Creator script (Quaranta, 2015, 2016).