Identifying Family Interrelationships in the World's Largest Census Microdata Collection

Sobek, Matthew
University of Minnesota, Minnesota Population Center
225 19th Ave South
Minneapolis, Minnesota 55406, United States of America
sobek@umn.edu

Kennedy, Sheela
University of Minnesota, Minnesota Population Center
225 19th Ave South
Minneapolis, Minnesota 55406, United States of America
kenne503@umn.edu

Census microdata describe the characteristics of individuals and give researchers the freedom to calculate their own measures of demographic and social phenomena. In most census datasets individuals are organized into households, and the relationships among individuals are known. This hierarchical structure gives the data much of its power. Researchers can combine the characteristics of related and co-resident persons to create a wide range of new variables and measures, such household structure, age of a person's spouse, or the number of own children present for each adult woman. Although the ability to create variables from multiple person records is essential for many analyses, it is difficult and error-prone. Censuses typically identify each person's relationship to a reference person in their household, but the relationships to other persons are often ambiguous and involve complex logic and data manipulation.

Consistent family interrelationship measures are especially critical for comparative research. This need is addressed by the IPUMS-International project, the world's largest collection of publicly available census microdata. The International Integrated Public Use Microdata Series consists of 325 million person records in 158 census samples from 55 countries (Minnesota Population Center 2010). Family relationship variables have been developed to produce a consistent set of links between immediate family members. By capitalizing on the hierarchical structure of the data, these variables give researchers the flexibility to define their own measures of household composition and to interrelate the characteristics of family members in complex ways (Ruggles 1995).

The IPUMS-International database is designed for comparative research. Variables are harmonized across countries, so all samples use consistent codes. Integrated documentation describes the comparability issues that cannot be adequately conveyed through variable codes and labels. All data are available at no charge through a web-based data extraction system that provides pooled extracts containing only the samples and variables requested by researchers. Researchers download the microdata and analyze it themselves on their desktop. Individuals are organized into households in most of the database, and family interrelationship variables have been created for all these samples. Moreover, 13 samples included a question on the census questionnaire that asked respondents to identify the location (the line number) of their spouse and parents. We use these census pointers to evaluate the IPUMS constructed family pointers.

Locator variables—“pointers”— identify the position of each person’s mother, father, or spouse, if one is present in the household. The pointers are the basis of all family interrelationship variables in IPUMS. Consider the 8-person household shown in Table 1. The census relationship-to-household-head variable describes a number of family interrelationships. We know the head and spouse are parents of the three children and that the head and spouse are married to one another. For other household members, additional variables must be used to infer relationships, including marital status, the number of children-ever-born, and

proximity to each other. The right-most columns in Table1 show the constructed IPUMS pointers. The variable SPLOC records the person number of each person's spouse or partner. In this example, the head and spouse "point" to each other (receiving SPLOC values 2 and 1 respectively). The variables MOMLOC and POPLOC provide the person number of each individual's parents—so the grandchild in position 7 points to his mother in position 6 and his father in position 5. When no spouse or no parents are identified, the pointer variables equal zero.

Table 1. Example of census household with constructed pointers

Person number	Relationship	Age	Sex	Marital status	Children ever born	SPLOC	MOMLOC	POPLOC
1	Head	73	Male	Married	n/a	2	0	0
2	Spouse	62	Female	Married	6	1	0	0
3	Child	38	Female	Single	1	0	2	1
4	Child	30	Female	Cohabiting	0	0	2	1
5	Child	32	Male	Married	n/a	6	2	1
6	Child-in-Law	30	Female	Married	1	5	0	0
7	Grandchild	6	Male	Single	n/a	0	6	5
8	Employee	16	Female	Cohabiting	Unknown	0	0	0

Because the same rules are applied across samples, households with similar characteristics in different countries or different years of the same country will receive the same distribution of constructed pointers. Moreover, the pointer variables will be identical for every researcher who downloads IPUMS data. Once SPLOC, MOMLOC, and POPLOC are created, additional family relationship variables are constructed, including the identification of subfamilies, the calculation of the number of children who are linked to a particular woman, and the number of families in a household. A feature of the IPUMS data extract system lets researchers attach the characteristics of parents and spouses as new variables on each person's record; thus they never have to use the pointers to perform that matching procedure in a statistical package.

The family presented in Table 1 is small, provides detailed relationship information, and requires only one decision—a relatively easy choice between the grandchild's two possible mothers. Producing family pointers becomes substantially more difficult when the relationship pairings are more ambiguous, when parental absence or adoption occurs commonly, or when there are multiple potential spouses and parents.

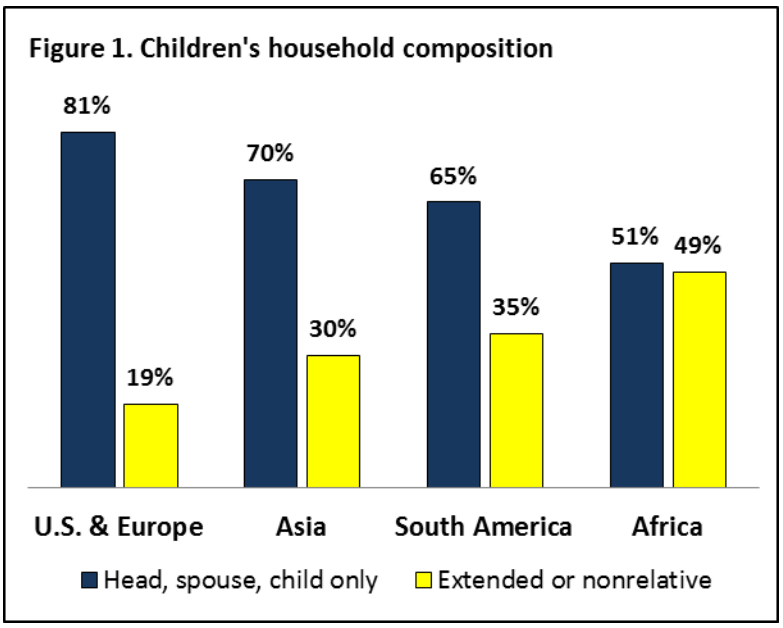
IPUMS-International pointer design

The IPUMS pointers are rule-based, evaluating individual pairings based on relationship to head, age, marital status, fertility (when available), and proximity in the household. The most important factor governing the development of international family interrelationships is the varying size and complexity of households around the world. Polygamy, prevalence of non-marital fertility, and the common presence of extended family members and nonrelatives make family interrelationships more uncertain for a higher proportion of individuals in some samples. To illustrate this diversity, Figure 1 presents data on the regional variation in the composition of children's households. Only half of children in the IPUMS African samples live in a household containing only the head, at most one spouse, and children of the head, compared to over 80 percent of children in the U.S. and Europe.

Also important is variation in the data available to construct the pointers (IPUMS website). Many samples, for instance, do not distinguish parents from parents-in-law or children from children-in-law, or they group grandchildren with other relatives. Data on children ever born or surviving—information that takes on considerable importance when relationship pairings are weak or when there are multiple potential parents—is sometimes unavailable.

The relative position of individuals within households is a strong indicator of family interrelationships in many samples. Censuses commonly instruct enumerators to list household members in meaningful

groupings. The IPUMS linking program capitalizes on this common feature of censuses by searching first for adjacent persons or chains of persons. However, the meaningfulness of household order for family interrelationships varies across samples.



Although some customization is necessary to handle particular situations, the same core conditions and basic linking methods are applied across all samples. Each household is evaluated individually. For each of the pointers, the program makes a series of passes looking for a spouse or parent. The strongest possible criteria are applied first to identify the most iron-clad links. Persons who are linked are removed from consideration by the subsequent, weaker passes that use more ambiguous criteria.

The simplest of the family interrelationship pointers is the location-of-spouse variable (SPLOC) that identifies the person number within the household of each individual's co-resident spouse or partner. The spouse pointer is easier to construct than the parental pointers because we know the person's current marital status, spouses generally reside together, and most people only have one spouse. Nevertheless, there are various complications, and the quality of the links varies across samples because of differences among the key variables and in the organization of persons within households.

The basic algorithm for SPLOC restricts the allowable pairings based on age, sex, marital status, and relationship to the household reference person. A linked couple must be of opposite sex and both persons must be age 12 or older. Links can only be made between persons in the same subfamily in the small number of samples that report such subunits. Both persons in a couple must indicate that they are in a marital or consensual union.

Starting with the first record in a household, each person is evaluated using the strongest possible criteria to locate a probable spouse. The strongest criteria involve explicit relationship combinations such as head-to-spouse and parent-to-parent. Subsequent passes use progressively weaker rules to make links—generally based on more ambiguous relationship pairings. At the moment a person is linked they and their spouse are removed from further consideration, thus the order in which the passes are executed is determinative. In most households there is only one possible married couple, and the accuracy of the link is nearly certain. Where there are multiple equally valid potential spouse candidates, the persons' proximity within the household roster is used to choose among them. A separate variable indicates the specific set of conditions under which each link was made.

Limited information on cohabitation in some samples poses the most serious comparability issue for the spouse pointer. Out of 115 samples, nine identify unmarried partners of household heads only in the relationship variable. Partners of persons other than the household head therefore cannot be identified; however, since these samples are exclusively from developed countries with relatively simple household structures, the great majority of consensual unions are undoubtedly recorded. Finally, polygamy poses a technical complication for the spouse identifier. Where polygamy is indicated, multiple females can link to one man; but he in turn can link to only the most proximate spouse, because the spouse pointer variable can only record a single person number.

Links between children and parents occur after the creation of spousal links. Unlike the spouse pointer, there is no variable comparable to marital status that consistently indicates a person’s eligibility to receive a

parent link. Consequently, all persons are considered eligible to be "children." All adults are eligible to be parents, although fertility plays an important role in evaluating parent-child pairings. Fertility data cannot be determinative for two reasons: first, because our parent pointers are designed to identify both biological and social parents; and because roughly one quarter of IPUMS samples contain no fertility data, while others limit this information to married or reproductive age women.

Like the spouse pointer, the parental linking algorithm works sequentially downwards through a household. Each person is evaluated in turn as a potential "child," and the program searches the household for a probable mother or father, based on relative ages, relationship to head, parent's marital status, mother's fertility, and proximity. The specific criteria used to evaluate a possible match depend on the child's relationship to the household head and fall under five broad rules based around the allowable pairings.

Most parental links are unambiguous—like a link between the household head and a child of household head—and 94% of all parent links fall under Rule 1, the strongest rule. Other links are less certain, such as links between children and grandchildren, or between nonrelatives of the head. As links become weaker, the criteria for matching become more stringent. For instance, adjacency is required for links involving nonrelatives of the head while additional age and fertility requirements are implemented when linking children to polygamous spouses. The algorithm searches for both fathers and mothers simultaneously, but within a given strength test links to potential mothers are evaluated before links to potential fathers.

Once a link is made several variables are automatically generated. The first is a rule variable (PARRULE), which describes the specific conditions under which the parent pointers were produced. We also produce stepmother and stepfather variables to identify links that are definitely or probably not biological: including links to explicitly-identified adopted and stepchildren, links in excess of a woman's known fertility, and links that fall outside reproductive age ranges.

A primary concern in the development of the pointers was to prevent all children in complex households from linking to a single parent when there were multiple legitimate candidates. This is particularly salient where the ordering of the persons within households makes proximity an ineffective linking criterion, for instance when all grandchildren are grouped together instead directly following their parents. To address the problem, we rely heavily on reported children ever born and children surviving to determine how many children should link to a particular woman and, by extension, to her spouse or partner. We refer to this as the "child cap" for a parent or couple. In some contexts, the linking algorithm allows the cap to be exceeded, but only after other potential parents have received their share of eligible children.

Unfortunately, some censuses do not collect women's childbearing data and virtually no countries collect data for men. In these instances where we could not use empirical data to "cap" links to a potential parent, we needed some way to apportion children among potential parents. To do this, we calculate a child cap which our algorithm uses in place of known fertility. The caps are based on the five rules for linking children to parents. Children are allocated among parents in proportion to the total number of children eligible to link to each parent under a particular rule. In addition, the caps are designed to increase the probability that we link to ever-married compared to never-married women. In households with a small number of children to be linked and many potential parents, the estimated cap will tend to divide the children evenly among all potential parents.

In order to assess the performance of the calculated child caps, we produced two sets of pointers for the samples with fertility data: one set using calculated caps and one using known fertility. For 98.5 percent of children under age 18, the two sets of maternal pointers agreed completely. The pointers constructed without fertility data do, however, slightly overestimate early and non-marital fertility.

Comparison to census pointers

Thirteen international censuses directly asked respondents for the line number on the census form of their mother, father or spouse. These links were used for guidance during the development of the IPUMS pointers, and they provide a means to evaluate the final product. Although the samples over-represent

Europe, they are nevertheless diverse, including both developed and developing countries, and have temporal depth.

The rate of disagreement between the IPUMS pointers and the corresponding pointers from the censuses is presented in Table 2. Overall, the spouse pointers agree 99.5% of the time, and the parental pointers more than 98.7%. The denominator for the mother and father statistics is all persons, because even adults are at risk of residing with parents. If one considers parental links only to persons under age 18, the rate of disagreement roughly doubles. Still, the absolute level of agreement is over 97%.

Table 2. Disagreement between IPUMS and Census Pointers (%)

		All persons		Age < 18	
Census	Spouse	Mother	Father	Mother	Father
Armenia 2001	1.29	1.09		2.62	
Belarus 1999	0.16	0.28		0.43	
Brazil 1991		0.46		1.30	
Portugal 1981	0.32	1.11	0.39	1.06	0.46
Portugal 1991	0.15	1.92	0.63	1.27	0.61
Portugal 2001	0.23	0.61	0.31	1.39	0.91
Romania 1977	0.36	0.43	0.21	0.62	0.44
Romania 1992	0.54	0.36	0.29	1.09	0.93
Romania 2002	0.11	0.20	0.17	0.69	0.63
South Africa 2001	1.21	4.87	1.88	9.96	4.08
South Africa 2007	0.83	3.89	1.28	8.88	2.92
Spain 1991	0.10				
Spain 2001	0.20	0.30	0.23	0.55	0.51
TOTAL	0.46	1.29	0.60	2.49	1.28

The rate of disagreement varies across samples due to a variety of factors. The IPUMS linking algorithm is designed to be sensitive to the reporting order of persons within households, but some samples are less well ordered than others because of differing enumeration practices or post-enumeration data processing. Samples also vary in their rate of data errors in substantive variables. The category detail in the key variables also differs, producing more ambiguous situations for the pointer algorithm in some samples.

The linking success rate is also affected by the underlying social reality reflected in the data. Some situations and living arrangements are inherently more difficult for the pointer program to manage. Basically, the more complex the household structure, the more chance there is to make an error. At the sample level, the correlation between the discrepancy rate and the proportion of persons living in extended households is .89 for spouse links, and .86 and .83 for mother and father links. The samples with census pointers have smaller households on average than the full IPUMS database: 4.7 persons versus 5.4 persons per household. They also have fewer persons living in extended families: 29.8% compared to 33.5%. It is therefore likely that the constructed pointer variables for IPUMS as a whole are somewhat less accurate than the average rates suggested by Table 2.

A majority of mismatches between IPUMS and the census pointers involve situations where the census did not identify a parent or spouse, yet IPUMS linked to someone who met the necessary criteria. Such errors of commission are to some degree unavoidable. If there is any plausible parent or spouse in the household, the IPUMS program will link to them. For spouses, there is usually no way of knowing the correct partner is absent. For mothers and fathers there is sometimes supporting evidence on fertility history or parental mortality that suggests the biological parent is absent. But because the IPUMS pointers are intended to encompass social parentage—step and adopted children—the linking tends to be generous, even to the point

of exceeding the known number of children a woman has borne.

Globally, less than 2 percent of persons live in a situation where there is more than one potential mother, father or spouse to whom they could conceivably link. Apart from the issue of absent persons, these complex situations pose the greatest challenge for the linking program; and in some African and Asian countries they can be several times more common than the world average. Where there is a choice to make, IPUMS points to a different spouse 11% and a different father 15% of the time. Mothers have a 26% discrepancy rate, driven substantially by South Africa, where over one-third of the links are different. The mean rate for the other 10 samples is 14% for mothers. The error rate for South Africa may be indicative of factors that are likely to obtain elsewhere in Africa, but it could be an idiosyncrasy of the data collection practices in one country. In any case, the conclusion is that the pointers produce relatively high error rates in a small subset of households.

Overall, the analysis suggests that individual level disagreements between the IPUMS and census pointers occur rarely, except in complex households. These disagreements, however, may balance out in the aggregate, if the IPUMS links are representative overall of actual family relationships. For instance, although we may link a grandchild to the wrong parent, it is likely that the true parent, a sibling, will be similar in age, education, or even marital status. Aggregate statistics with respect to age difference between spouses and between parents and children, and in the propensity of children to live with one or both parents are all quite similar using census or IPUMS pointers.

Discussion

The census pointers, direct reports of spouse and parent location collected during enumeration, provide valuable information on the strengths and limitations of the IPUMS pointers. Across samples with empirical census pointers, the IPUMS and census pointers are in close agreement, although disagreement rates are higher for individual countries. The agreement rate falls notably when there are multiple spouse or parent candidates, but fewer than one-in-fifty persons face such a choice. The characteristics of spouse and parent-child pairings produced by the IPUMS pointers resemble closely those of census pairings. Estimates of the age-differences between spouses are virtually indistinguishable in all samples, while estimates of early motherhood are only slightly higher using the IPUMS pointers. The IPUMS estimates of residence with a lone mother or with an unmarried motherhood are also higher, on average, than census pointer estimates. Additional analyses indicate pointers constructed in samples without fertility data will slightly overestimate early and unmarried motherhood. But even in the most challenging situations, limited data or complex household structure, the IPUMS pointers perform well.

The major factors that affect our ability to correctly identify spouse and parent-child pairs include: the meaningfulness of household order; the availability of detailed relationship categories, such as grandchildren or in-laws; the availability of childbearing data; and the size and complexity of households structures. Differences in these factors will influence the comparability of the pointers across countries and within countries over time. In general, we expect the impact to be quite small, but researchers must evaluate the likely effect for their individual projects. For example, in samples without fertility data, overall estimates of non-marital fertility are likely to track closely the true population levels, but a non-trivial minority of the children who receive links to unmarried mothers may do so erroneously. To assist researchers, the IPUMS project documents the differences among samples in the available raw materials for the construction of these links. This allows researchers to make informed decisions when their object of study might be especially susceptible to particular limitations in the underlying data.

REFERENCES

Minnesota Population Center. 2010. Integrated Public Use Microdata Series, International: Version 6.0. Minneapolis: University of Minnesota.

Ruggles, Steven. 1995. “Family Interrelationships.” Historical Methods 28:52-58.