The Development of Family Interrelationship Measures for International Census Data

Sheela Kennedy
Matthew Sobek
Minnesota Population Center

DRAFT – please do not cite

September 18, 2009

The Development of Family Interrelationship Measures for International Census Data

Population microdata are often organized into households, but taking full advantage of the hierarchical structure of the data requires identifying the specific family interrelationships among household members. Household relationships are expressed relative to a single reference person and are often ambiguous for persons outside the nuclear family. The International Integrated Public Use Microdata Series (IPUMS-International) has developed consistent "pointer" variables for 115 census samples, identifying each person's mother, father and spouse. The database, with over 279 million person records, is freely available for downloading by researchers. Some countries collected pointer information as actual census questions, and the IPUMS pointers agree with these data 98% of the time. Nevertheless, some household situations are more problematic than others, and there is variation in the source materials across countries. Although imperfect, by using these common tools researchers remove the possibility that differing results across studies are an artifact of different linking procedures.

Introduction

Census microdata are among the most widely used sources in population research. 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, and can analyze their effects at the individual level. Constructed variables might include the age of a person's spouse, the school attendance of a father's children, or the number of own children present for each adult woman in the household.

The ability to create variables from multiple person records is essential for many analyses, but it is an inherently difficult task. Censuses invariably identify each person's relationship to a reference person in their household, but the relationships to other persons are often ambiguous. Some people are grouped into residual "other relative" and "non-relative" categories, and even persons with a specified relationship like "grandchild" might have more than one potential mother or father. Adequately determining family interrelationships requires using a number of variables in combination and considering factors like persons' proximity within the household roster. Many researchers are unable to carry out these methods in a statistical package. But even those with sufficient skill to make such links will invariably use differing methods, because of the many decisions embedded in such techniques. Consequently, there will always be uncertainty about the extent to which differing results between researchers are artifacts of the linking process.

This paper describes and evaluates the development of consistent family interrelationship information for the world's largest collection of publicly accessible census microdata. The

International Integrated Public Use Microdata Series (IPUMS-International) consists of 279 million person records in 130 census samples from 44 countries. Family relationship variables have been developed for 115 of these samples.[1] These "pointers" are designed to produce a consistent, but flexible set of links between immediate family members. By capitalizing on the hierarchical structure of the data, the pointers give researchers the flexibility to define their own measures of family and household composition and to interrelate the characteristics of family members in complex ways. All users of the IPUMS database have access to these variables. The basic task of making a new variable by attaching a characteristic of one person to another -- age of mother, for example -- can be carried out by the web-based data extraction system. Thus researchers need not be familiar with the mechanics of sorting and matching to take advantage of these powerful tools.[2]

Background

IPUMS International Project

The IPUMS International project was developed at the Minnesota Population Center with the goal of cataloging, preserving, harmonizing and disseminating international census microdata and documentation (Hall, McCaa, and Thorvaldsen 2000). Housed at the Minnesota Population Center, the 2009 version of the data series includes data from 44 countries, and the project has agreements with an additional 38 countries to make their data available in the future. For most countries, data are available for multiple census years.

The database is designed for comparative research. Variables are harmonized across countries, so all samples use consistent codes. No information is lost. For more complex variables, the first digit or two are comparable across samples, while trailing digits retain information unique to particular samples. Integrated documentation describes the comparability issues that cannot be adequately conveyed through variable labeling and coding schemes. 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 115 samples from 42 countries, and family interrelationship variables have been created for these samples. The full list of these samples is shown in Appendix 1. In addition, 13 samples included a question on the census enumeration form that asks respondents to identify the location (the line number) of each person's spouse and parents. We use these census pointers to evaluate the IPUMS constructed family pointers.

Family Interrelationship variables

Our primary goal was to produce a series of family locator variables or "pointers"—variables that identify each person's mother, father, or spouse, if one is present in the household. Consider the 8-person household shown in Table 1. The 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,

[1] Linking variables could not be constructed for some datasets because the person records were not organized into households or because they lacked a critical variable for making the links.

[2] The IPUMS-International data series is continually growing and evolving. The discussion in this paper pertains to the database and its constructed variables as of fall 2009 (Minnesota Population Center 2009).

additional variables must be used to infer relationships, including marital status, the number of children-ever-born, and proximity to each other in the household. The female child-in-law is almost certainly married to the preceding child, because both share the same marital status and because there are no other male children to whom she could be married. The grandchild, however, could be the son of the female child in position 3 (a single mother of one child). More likely, however, he is the son of the adjacent married couple (the child and child-in-law listed directly above him in the household).

Table 1. Example of census household

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

Rather than forcing researchers to work through the complex logic to define these family interrelationships, IPUMS constructs the necessary variables using the same program for all samples. These "pointer" variables give the person number in the household of each individual's mother, father and spouse. Table 2 shows the constructed pointers for the same household described above.

Table 2. Example of census household with constructed pointers

Person				**Marital**	**Children**			
number	**Relationship**	**Age**	**Sex**	**status**	**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

The variable SPLOC contains the person number of each person's spouse or partner. In this example, the head and spouse "point" to each other (receiving SPLOC = 2 and 1 respectively). The variables MOMLOC and POPLOC provide the person number of an 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 are given the value zero.

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 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 Tables 1 and 2 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. The challenge we faced was to develop a consistent set of family relationships that could be applied to countries that differ greatly in family and household structure and in the detail and quality of data.

Origins of matching procedures

The origins of family interrelationship construction can be found in the "own-child" method of fertility measurement. First developed in the early 1960s and refined in later years, the own-child method estimates fertility using census data when birth registration data are incomplete or unavailable (Grabill and Cho 1965; Retherford and Cho 1978; Retherford, Cho, and Kim 1984; Luther and Cho 1988). Within each census household children are matched to mothers, using an algorithm that incorporates demographic data usually collected during census enumeration: relationship to household head, age, marital status, and the number of surviving children, when available.[3] Reverse survival methods are then used to estimate the number of children born in a particular year, as well as the number of women by age. From this, single-year age-specific fertility rates can be calculated for periods up to 15 years prior to census enumeration.

Own-child methods have been used widely to estimate international and historical fertility levels. Researchers continue to use these methods when birth registration data are not available, often to provide estimates of historical trends in fertility (Retherford et al. 2005; Hacker 2003; Zuberi and Sibanda 1999). Comparisons have found that own-child matches yield similar population level fertility estimates as direct reports of mother-child relationships, even in a sample with complex families, high rates of adoption, and a high rate of mismatches (Levin and Retherford 1982; Cho et al. 1986). Although individual-level errors tend to cancel out when aggregated, errors rates can be high at the extreme ends of the reproductive age range. More complex matching procedures have since been developed, but have not been implemented widely (Zuberi and Sibanda 1999; Strong et al. 1989).

[3] Examples of matching programs are included in Cho, Retherford, and Choe (1986).

With the 1995 release of integrated microdata files for eleven U.S. censuses, IPUMS-USA advanced the process of identifying family relationships (Ruggles 1995). Family interrelationship variables were reconceived as a set of multi-purpose tools made available to researchers in public use samples. IPUMS-USA provided additional family pointers not included in own-child methods (links between spouses and between children and their fathers) and constructed additional family and household descriptors. The algorithm had to be flexible enough to deal with differing variable availability and changing category detail across census years. Each pointer variable was accompanied by a rule variable describing the criteria used to assign the spouse or parent link. The resulting family interrelationship variables have allowed researchers to study a variety of topics, no longer limited to fertility. These topics include historical estimates of family and household composition and studies of family structure and child wellbeing (Ruggles and Brower 2003; Moehling 2004, 2007; Short, Goldscheider, and Torr 2006; McGarry and Schoeni 2000; Lichter, Qian, and Crowley 2008).

The IPUMS parental pointers were deliberately conceived to include social parents, not simply biological ones. For one thing, it was not always possible to distinguish between the two, because of differing category and variable availability among samples. More importantly, for many research purposes, links identifying social and economic units are more desirable than ones limited to biological connections. Because biological links are sometimes necessary, IPUMS provides supplemental variables that identify whether a given mother or father is likely a step-parent.

IPUMS International Pointer Design

We initially tried adapting the IPUMS-USA algorithms for the international IPUMS project, but soon determined that the U.S. model could offer only rough guidance. The international samples simply had too much variation: in the reporting order of the enumerated persons, in the categories of the relationship-to-head variable, in the types of marital statuses, and in the quality of the data. The U.S. database's focus on social parentage, rather than strictly biological links, was retained in the international data series.

Perhaps the most important factor governing the development of international family interrelationships is the increased size and complexity of households in the IPUMS samples. Links between the spouse of the head and a child of the head, which are unambiguous in the U.S., are less certain in samples with polygamy. Rising rates of non-marital fertility mean that matching procedures cannot exclude never-married women. Likewise, the more common presence of extended family members and nonrelatives, makes family interrelationships more uncertain for a higher proportion of individuals. To illustrate this international diversity, Figure 1 presents the data on the international 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. To compound the difficulty, many of the same samples with large numbers of complex households have relatively high rates of data errors in key variables like age, sex, and relationship.

Also important is variation in the data available to construct the pointers[4]. 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—are often unavailable. Finally, the ordering of persons within households is often not as meaningful in the international data as in the samples comprising IPUMS-USA.

The IPUMS-International project emphasizes consistency across samples in the design of family interrelationship variables. 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.

SPLOC

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 (see Appendix 2). The strongest criteria involve explicit relationship combinations such as head-to-spouse, parent-to-parent, etc. 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, so 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.

The biggest challenge in developing the spouse pointer was determining which relationship-to-head categories could link to one another. Theoretically the allowable pairings should be a

[4] For information on sample availability of key data used in pointer construction (e.g. relationship categories and fertility) see https://international.ipums.org/international/parrule_table.shtml.

straightforward inference from the relationship labels: spouse-to-head, child-to-child-in-law, etc. But matters are complicated considerably by differences in category availability, terminological slippage across samples, and data inconsistencies. For example, in some samples the "sibling" category may have included large numbers of siblings-in-law; or "spouse" might mean the wife of any household member rather than exclusively the wife of the head.

We required a method to systematically uncover these irregularities. Accordingly, for every sample we calculated the number of additional couples that would be created if we allowed any given pair of relationships to link. This "matchmaker" program produced a list of possible pairings in each sample that warranted further examination: those that involved a non-trivial proportion of the total married population and whose allowance would substantially reduce the spouse-absent rate of one of the involved relationship categories. Our analysis led to refinement of the basic list of acceptable pairings and to sample-specific customizations, such as allowing child-to-child links in samples where "child" apparently includes children-in-law.

The matchmaker method also exposed complications related to the reporting of marital status and cohabitation across samples. Close inspection revealed that obvious couples commonly gave different "in union" responses: for example a household head said he was legally married but his spouse reported being in a consensual union. It therefore proved necessary to globally allow mismatched statuses as long as both persons reported some kind of union and had appropriate relationship information. In selected instances relationship information can even override marital status: spouses can link to heads even if only one of them claims to be in a union, and unmarried partners can link to heads regardless of their marital statuses. We also uncovered widespread uncertainty among census respondents about whether the consensual partners of heads and family members should be called relatives or non-relatives. By definition, non-relatives should never be linked to relatives; but where consensual unions are concerned, that fundamental divide cannot be maintained. Consequently, after all other passes have been made "other relatives" and non-relatives can link to heads or any other family member, as long as both parties report being in consensual unions.

Polygamy poses a technical complication for the spouse identifier. Where polygamy was 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 only records a single person number. In samples in which only men are identified as polygamous in the marital status variable, multiple women can link to a polygamous man as long as the women are in a marital union of some kind. Finally, some samples do not identify polygamous unions, although polygamy was widely practiced. We allow multiple female spouses to link to heads in those censuses. Polygamous unions not involving the head and spouse cannot be identified, but such unions are much less common, as we can determine from the samples that do identify polygamous unions.

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. More troubling are the 14 samples from censuses whose questionnaires specified only legal unions were to be reported. In some of these societies, consensual unions were probably rare, and in others it is possible that substantial numbers of de facto marriages were reported regardless of what the census instructions may have stipulated. Analysis of European countries that changed the legal-status requirement between censuses suggests the instructions had little effect on the overall distribution of responses; but there is no way to tell with certainty, and some affected countries lack data with which to make comparisons.

MOMLOC/POPLOC

Links between children and probable parents occur after the creation of spousal links. Unlike marital status, no comparable variable exists in all samples that could be used to determine a person's eligibility to receive a parent link.[5] Consequently, all persons are considered eligible to receive parent links except for unrelated persons over age 15 and related persons over age 15 with an unspecified relationship to the household head.[6] All adults are eligible to be parents, although fertility plays a critical role in evaluating parent-child pairings. This is necessary for two reasons: first, because our parent pointers are designed to identify both biological and social parents; and second, because over 25% of IPUMSI samples contain no fertility data, while others limit this information to married or reproductive age women.

Like SPLOC, the MOMLOC/POPLOC algorithm works sequentially downwards through a household. We first identify a "child" and then search the household for a probable mother or father, based on relative ages, relationships to head, marital status, fertility, and proximity. The specific criteria used to evaluate a potential match depend on the child's relationship to the household head (RELATE) and fall under five broad rules. Appendix 3 describes the allowable pairings in each rule, by relationship, age differences, fertility, and household position. Most 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, for instance links between children and grandchildren, or between nonrelatives of the head. As links become weaker, our criteria for matching become more stringent. For instance, never-married non-cohabiting men are eligible to be fathers only when the relationship-pairing is unambiguous. Although the algorithm searches for both fathers and mothers simultaneously, within a given strength test links to potential mothers are evaluated before links to potential fathers. As soon as a link is made, either to a mother or to a father, a second link is automatically generated to that person's spouse or partner, and no additional attempts are made to find parents for that individual.

Once a link is made several variables are automatically generated. The first is PARRULE, which includes the specific rule under which MOMLOC and POPLOC were produced. We also

[5] For instance, data on mother and father mortality (MORTMOM and MORTPOP) are available for only 16 samples. These variables are also not comparable to marital status, because children with a deceased parent may live with a stepparent, while some children with living parents may live apart from their parents. MORTMOM and MORTPOP data are included in the construction of variables indication likely stepparent relationships (STEPMOM and STEPPOP).

[6] Empirical evidence from samples with census pointers indicated that among persons 15 and older, only about 1 percent of nonrelatives and 5 percent of relatives with an unspecified relationship to the head lived with their parents. We concluded that given the low numbers of matches that should be made and the ambiguity of these relationship categories, we could not successfully construct pointers for these individuals.

produce STEPMOM and STEPPOP, variables which identify links that are definitely or probably not biological links: 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 (see Table 3). Using STEPMOM, researchers interested in fertility can select only those mother-child links which probably reflect biological relationships. We should note that there are many adopted and step parents who cannot be identified with information available in the censuses. Therefore, STEPMOM and STEPPOP will always under-represent their actual number in the population.

Table 3. STEPMOM values

0 = Biological mother or no mother present
1 = Mother has no children born or surviving
2 = Child reports mother is deceased
3 = Explicitly identified step relationship
4 = Mother reports no children in the home
5 = Age difference implausible
6 = Child exceeds known fertility of mother

We used several approaches to guide the development of the MOMLOC and POPLOC algorithm. Whenever possible we relied on empirical evidence drawn from IPUMS samples. For example, information on samples that distinguish between children and children-in-law was used to develop procedures to apply to samples that combined these categories. We also used information from the small number of samples that collected data on parent's location in the household as part of census enumeration. If a particular specification produced links with a high rate of disagreement with the census pointers (described more below) we rejected the rule. For instance, our final algorithm disallows links to never-married non-cohabiting men as potential fathers unless the relationships were unambiguous, because these links were invariably wrong.

When the above methods were not possible, we implemented a procedure for voting on changes to the program. Each time we modified the algorithm we selected a random sample of about 500 households in which MOMLOC or POPLOC had changed as a result of the modification. These households were divided among several analysts who examined each household by hand, comparing the new pointers to the previous version and scoring each change as improved, worsened, or indeterminate. A change was accepted only if all analysts agreed that the modification resulted in a noticeable improvement.

A primary concern guiding 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.[7] Whenever possible, we relied heavily on reported children ever born (CHBORN) and children surviving (CHSURV) to determine how many children should link to a particular woman and 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 typically only after other potential parents have received their share of eligible children. Thus, the child cap plays a powerful role in the allocation of children.

[7] Roughly 98 percent of persons under age 18 had at most one person who qualified as a mother. In 12 samples more than 5 percent of children had 2 or more potential mothers, including 3 samples exceeding 10 percent of 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 calculated a child cap for potential parents which our algorithm uses in place of known fertility. The caps are based on the five rules for linking children and parents. Children are allocated among parents in proportion to the total number of children eligible to link to each parent under a particular rule.[8] In addition, the caps are designed to increase the probability that we link to ever-married potential mothers compared to never-married women, in recognition of the higher fertility of married women.[9] Never-married, non-cohabiting men do not link to children unless the relationship pairing is unambiguous (e.g. head and child).

Calculated caps are also used in instances where fertility data is available for only some women in the household. In these instances, caps are calculated only for persons with unknown fertility and take into account observed fertility of others household residents. In addition, our algorithm prioritizes links to persons with observed fertility over links to persons with a calculated child cap.

These caps play a critical role in determining whether a child should link to a particular parent, except when relationships are unambiguous. When a child links to a parent, the caps of the linked parent and their non-polygamous spouse/partner are reduced. Once a potential parent's cap is filled, we search for alternative parents with an available cap. 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 (for instance, everyone links to one child), even when the household order suggests an uneven distribution is more accurate. We concluded that this was preferable to a no-cap situation, in which unreasonably large numbers of children would link to just one mother.

Research with Census Pointers

It is difficult to assess the quality of the constructed family links without having some basis for comparison. Fortunately, a number of 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.

[8] To calculate the cap, we first count a woman's "potential children", or the total number of children who meet the basic relationship and age requirements for a mother-child match. A 3-year old grandchild qualifies as a potential child of a 47-year old female child, but a 2-year old grandchild does not. Next, we calculate each woman's share of children as the ratio of her potential children to the sum of children who could potentially link under a particular rule. For instance if we calculated a child cap under Rule 2 (child-grandchild matches), we would divide an individual child's potential matches by the sum of all potential matches between children and grandchildren. This proportion is then multiplied by the total number of children available to match under that rule (for instance the total number of grandchildren in the household). Caps for men are calculated separately, but follow the same logic.

[9] Calculations for ever-married women exclude the potential children of never-married women when calculating the denominator. In essence, we divide all available children between the ever-married potential mothers in a household ensuring that children will be more likely to link to the married women. Calculations for never-married women include the potential children of married potential mothers; as a result, never-married women have a reduced, but non-zero, probability of linking to children compared to ever-married women.

Thirteen out of 115 IPUMS samples contain census variables indicating the location of a spouse or parent. The 13 samples are not perfectly representative of the complete database. Over half the samples are from Europe, with only one sample from Latin America and two samples from one African country.

The rate of disagreement between the IPUMS pointers and the corresponding pointers from the censuses is presented in Table 4. Overall, the spouse pointers agree 99.8% of the time, and the parental pointers more than 98.6%. The denominator for the mother and father statistics is all persons, because adults are at risk of residing with parents. If one considers parental links only to persons under age 18, the rate of disagreement increases roughly by a factor of 2.5. Still, the absolute level of agreement is very strong, at over 96%.

The rate of disagreement varies across samples due to a variety of factors. The reporting order of persons within households often conveys significant information about family relationships, and the IPUMS linking algorithm is designed to be sensitive to that information. 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 linking process, which compares information from multiple records, will also tend to uncover data inconsistencies that are not evident in person-level tabulations. The category detail in the key variables also differs, producing more ambiguous situations for the pointer variable code to navigate in some samples. Finally, the census linking variables are recorded as numbers referring to other lines on the census form. Numeric data collected in this manner are especially error prone, and close examination of the data suggests these variables to be among the noisiest in the census samples.

Table 4. Disagreement between IPUMS and Census Pointers (%)

		all persons		age < 18	
Census	Spouse	Mother	Father	Mother	Father
Armenia 2001	0.61	1.10		2.65	
Belarus 1999	0.08	0.38		0.67	
Brazil 1991		0.47		1.33	
Portugal 1981	0.16	1.11	0.45	1.06	0.67
Portugal 1991	0.08	1.92	0.69	1.27	0.80
Portugal 2001	0.12	0.62	0.35	1.40	1.07
Romania 1977	0.19	0.44	0.24	0.65	0.54
Romania 1992	0.36	0.37	0.34	1.12	1.10
Romania 2002	0.06	0.20	0.19	0.68	0.72
South Africa 2001	0.34	5.09	2.78	10.65	6.22
South Africa 2007	0.23	4.33	2.29	10.13	5.41
Spain 1991	0.14				
Spain 2001	0.10	0.30	0.23	0.55	0.49
TOTAL	0.19	1.34	0.82	3.15	2.28

Samples are weighted equally.

The denominator for the spouse column is persons in a union.

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 code 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 (for persons under age 18). The relationship is still strong but somewhat weaker between mean household size and error rates, suggesting that household complexity is the salient issue. The samples with census pointers have smaller households on average than the full IPUMS database: 4.71 persons versus 5.39 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 4.

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 simply no way of knowing the 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 the IPUMS parental links are intended to reflect social parentage as well: step and adopted children. Thus our linking tends to be generous, even to the point of exceeding the known number of children a woman has borne. IPUMS constructs variables identifying probable step mothers and fathers so researchers can exclude social parents from analyses requiring strictly biological links.

On the one hand, absent spouses and parents pose the most difficult situation for accurate linking. On the other hand, the lack of a spouse or parent sometimes indicates an error in the original census pointer data. Non-responses are indistinguishable in the microdata from an absent parent or spouse: both typically receive a code of zero in the data. In general, we would expect the non-response rates to be higher in less developed countries, but there is surprising variation in quality even among the developed nations. In any case, to the extent that there are missing data in the census pointers, the error rates in Table 4 are exaggerated.

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. In such households, how frequently do IPUMS and the census point to different persons? Where there is a choice to make, IPUMS points to a different spouse 11% and a different father 15% of the time.[10] 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.

The census pointers can also reveal the specific relationship categories that pose the most difficulty for the linking program. For the spouse links, the greatest number of errors involves children linking to the wrong child-in-law. The error rate is only about 3 percent, but these are large categories. It's not uncommon for there to be more than one possible child to whom a child-in-law might link, or for the spouse in such situations to be absent. The parental linking errors are dominated by grandchildren linking to children, with the great majority stemming from the South Africa samples. In South African households there are often strings of children followed by strings of grandchildren, and it is difficult to accurately assign people to the correct mother.

There are at least two additional factors that the census pointer data do not help us address. African and Muslim countries often allow polygamy, and polygamous households are especially challenging for determining family interrelationships. But among the samples with census pointer variables, only South Africa identifies polygamy, and it has too few such cases to generalize. The samples with census pointers also do not let us measure the effect of de facto versus de jure census practices. The de facto censuses enumerated persons where they happened to be at the moment of the census, while the de jure censuses recorded people at their usual place of residence. The de facto censuses should have higher rates of absent spouses and parents, but there is insufficient diversity in the samples with census pointers to explore the issue.

[10] To reduce the confounding effect posed by data errors, we exclude cases where either IPUMS or the census points at no one.

Discussion

This paper describes the development of location-of-spouse and location-of-parent variables for the IPUMS-International project. The IPUMS family interrelationship variables make possible comparative analysis of family and household structure for 115 census samples in 42 countries. The 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.

The census pointers offer the best available evidence on the strengths and limitations of the IPUMS pointers. Across samples with empirical census pointers, the IPUMS and census pointers are in close agreement: 99.8% for the spouse pointer and 98.6% for parent pointers, although disagreement rates are higher for individual countries. Most of the factors that complicate accurate linking are correlated at the country level. The less developed countries tend to have larger and more complex households, less consistent enumeration practices, and more data errors during data collection and processing. On the positive side of the ledger, most developing countries with large, complex households have fertility and often parental mortality data, which help significantly in making the parental links.

Despite large differences in census enumeration practices and in family and household structure, the IPUMS international pointers provide an important and sufficiently consistent set of tools for researchers wishing to identify family relationships. Although these data have only recently become available, studies using the IPUMS pointers already include research on intergenerational coresidence, changes in family size and children’s resources, measurement of single-parent families, and a study of children and international migration in the Philippines (Ruggles and Heggeness 2008; Lam and Marteleto 2008; Bryant 2008; Heggeness 2009). By identifying family relationships, the IPUMS pointers allow researchers to easily create their own measures of family and household composition or measures of family-level characteristics. The pointers are difficult for individual researchers to construct, and they would invariably do so in different ways. By using consistent pointers available to everyone, researchers can replicate each other's results and be certain they are measuring the same phenomena.

Bryant, John. 2008. “Children and International Migration.” Pp. 177-194 in *Situation Report on International Migration in East and South-East Asia*. Bangkok: International Organization for Migration, Regional Office for Southeast Asia.

Cho, Lee-Jay, Robert D. Retherford, and Minja Kim Choe. 1986. *The own-children method of fertility estimation*. East-West Center, East-West Population Institute.

Grabill, Wilson H., and Lee-Jay Cho. 1965. “Methodology for the measurement of current fertility from population data on young children.” *Demography* 50–73.

Hacker, J. D. 2003. “Rethinking the "early" decline of marital fertility in the United States.” *Demography* 40:605-620.

Hall, P. K, R. McCaa, and G. Thorvaldsen. 2000. *Handbook of international historical microdata for population research*. Minnesota Population Center.

Heggeness, Misty. 2009. “(Mis)Measuring Lone-Mother Families.” Detroit.

Lam, D., and L. Marteleto. 2008. “Stages of the Demographic Transition from a Child’s Perspective: Family Size, Cohort Size, and Children’s Resources.” *Population and development review* 34:225.

Levin, Michael J., and Robert D. Retherford. 1982. “The effect of alternative matching procedures on fertility estimates based on the own-children method..” Pp. 11–17 in *Asian and Pacific Census Forum*, vol. 8.

Lichter, Daniel T., Zhenchao Qian, and Martha L. Crowley. 2008. “Poverty and economic polarization among America’s minority and immigrant children.” Pp. 118-143 in *Handbook of families and poverty: Interdisciplinary perspectives*, edited by D. Russell Crane and Tim B. Heaton. New York: Sage.

Luther, Norman Y., and Lee-Jay Cho. 1988. “Reconstruction of birth histories from census and household survey data.” *Population Studies* 42:451–472.

McGarry, Kathleen, and Robert F. Schoeni. 2000. “Social security, economic growth, and the rise in elderly widows' independence in the twentieth century.” *Demography* 221–236.

Minnesota Population Center. 2009. *Integrated Public Use Microdata Series — International: Version 5.0*. Minneapolis: University of Minnesota.

Moehling, Carolyn M. 2004. “Family structure, school attendance, and child labor in the American South in 1900 and 1910.” *Explorations in Economic History* 41:73–100.

Moehling, Carolyn M. 2007. “The American Welfare System and Family Structure: An Historical Perspective.” *Journal of Human Resources* 42:117.

Retherford, Robert D., and Lee-Jay Cho. 1978. “Age-parity-specific birth rates and birth probabilities from census or survey data on own children.” *Population Studies* 567–581.

Retherford, Robert D., Lee-Jay Cho, and Nam-Il Kim. 1984. “Census-derived estimates of fertility by duration since first marriage in the Republic of Korea.” *Demography* 537–558.

Retherford, Robert D., Minja Kim Choe, Jiajian Chen, Li Xiru, and Cui Hongyan. 2005. “How far has fertility in China really declined?.” *Population and Development Review* 57–84.

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

Ruggles, Steven, and Susan Brower. 2003. “Measurement of household and family composition in the United States, 1850-2000.” *Population and Development Review* 29:73-+.

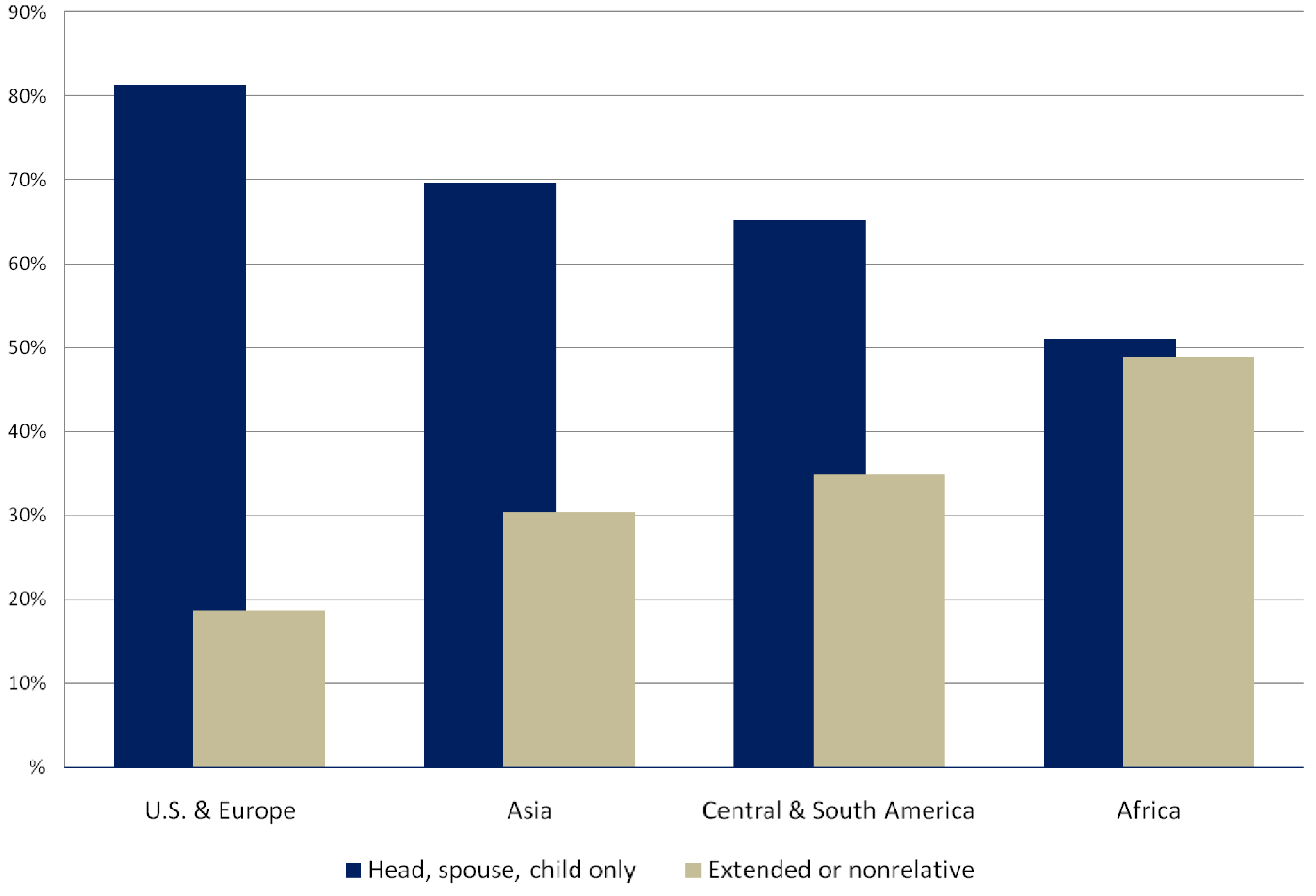
Ruggles, Steven, and Misty Heggeness. 2008. “Intergenerational coresidence in developing countries.” *Population and Development Review* 34:253-+.

Short, Susan E., Frances K. Goldscheider, and Berna M. Torr. 2006. “Less help for mother: the decline in coresidential female support for the mothers of young children, 1880-2000.” *Demography* 617–629.

Strong, Michael A. et al. 1989. *User's Guide: Public Use Sample, 1910 United States Census of Population.* Ann Arbor: University of Michigan, Inter University Consortium for Political and Social Research.

Zuberi, Tukufu, and Amson Sibanda. 1999. *Fertility Differentials in sub-Saharan Africa: Applying Own-Children Methods to African Censuses*. Philadelphia: University of Pennsylvania.

Figure 1. Regional differences in household composition, children ages 0-17



Appendix 1. IPUMS samples with family interrelationships variables

Country	Census years
Argentina	1970, 1980, 1991, 2001
Armenia	2001
Austria	1971, 1981, 1991, 2001
Belarus	1999
Bolivia	1976, 1992, 2001
Brazil	1960, 1970, 1980, 1991, 2000
Cambodia	1998
Chile	1970, 1982, 1992, 2002
China	1982, 1990
Colombia	1973, 1985, 1993, 2005
Costa Rica	1973, 1984, 2000
Ecuador	1974, 1982, 1990, 2001
Egypt	1996
France	1962, 1968, 1975, 1982, 1990, 1999
Ghana	2000
Greece	1971, 1981, 1991, 2001
Guinea	1983, 1996
Hungary	1970, 1980, 1990, 2001
India	1983, 1987, 1993, 1999
Iraq	1997
Israel	1972, 1983, 1995
Italy	2001
Jordan	2004
Kenya	1989, 1999
Kyrgyz Republic	1999
Malaysia	1970, 1980, 1991, 2000
Mexico	1970, 1990, 1995, 2000
Mongolia	1989, 2000
Palestine	1997
Panama	1960, 1970, 1980, 1990, 2000
Philippines	1990, 1995, 2000
Portugal	1981, 1991, 2001
Romania	1977, 1992, 2002
Rwanda	1991, 2002
Slovenia	2002
South Africa	1996, 2001, 2007
Spain	1991, 2001
Uganda	1991, 2002
United Kingdom	1991
United States	1960, 1970, 1980, 1990, 2000, 2005
Venezuela	1971, 1981, 1990, 2001
Vietnam	1989, 1999

Appendix 2. Rules for construction of SPLOC

The detailed order of the linking passes through each household is as follows:

1. Strong relationship pairing, both persons in some kind of union
2. Weak relationship pairing, exact marital status match
3. Weak relationship pairing, differing marital statuses
4. Non-relative link to any relative, both persons in consensual unions
5. Head link to non-relative, both persons in consensual unions
6. Head link to spouse, one person consensual union and the other not in a union
7. Sample-specific rules (child link to child), exact marital status match

"Strong" relationship pairings are links between specified relationships, like child-to-child-in-law, head-to-spouse, or aunt-to-uncle. "Weak" relationship pairings include at least one non-specific or ambiguous category, such as other relative or non-relative. Within each set of conditions above, first adjacent then nonadjacent persons are considered as potential spouses: specifically, preceding adjacent, following adjacent, preceding non-adjacent, and following non-adjacent persons. Where relationship pairings are ambiguous, the female can be no more than 20 years older than the male, and the male no more than 35 years older than the female (rules 2-5 and 7).

Appendix 3. Rules for construction of MOMLOC and POPLOC

			Age	CHBORN	Require	
Rule	**Child's relationship to head**	**Parent's relationship to head**	**difference**	**limits**	**adjacency**	**Notes**
Rule 1: Links involving Head and Spouse						
	Child	Head, spouse, unmarried partner	10-69	no	no	
	Child	Spouse/partner of polygamous head	10-54	weak	no	1
	Child/child-in-law	Head, spouse, unmarried partner	10-69	no	no	2
	Child/grandchild	Head, spouse, unmarried partner	10-44	no	no	3
	Head, sibling	Parent, parent/parent-in-law, parent/grandparent	10-69	no	no	
	Spouse, sibling-in-law	Parent-in-law	10-69	no	no	
	Sibling/sibling-in-law	All parent categories	10-69	no	no	
Rule 2: Links between grandchildren and children						
	Grandchild	Child, child/child-in-law	15-44	yes	no	4
Rule 3: Links between other specified relatives						
	Nephew/niece	Sibling, sibling/sibling-in-law	15-44	weak	no	4
	Nephew-in-law/niece-in-law	Sibling-in-law, sibling/sibling-in-law	15-44	weak	no	4
	Grandchild, great-grandchild	Grandchild	15-44	weak	no	4
	Cousin	Aunt/uncle	15-44	weak	no	4
Rule 4: Links involving other unspecified relatives and other relatives/non-relatives						
	Head	Other relative	>=20	strict	no	5, 6
	Other relative, other rel/non-rel	Child	15-44	strict	no	5, 7
	Other relative, other rel/non-rel	Unmarried partner	15-44	strict	no	5
	Other relative, other rel/non-rel	Other relative	15-44	strict	no	5
	Other relative, other rel/non-rel	Grandchild	15-44	strict	no	5
	Other relative, other rel/non-rel	Sibling, sibling-in-law	15-44	strict	no	5
	Other relative, other rel/non-rel	Other relative/non-relative	15-44	strict	no	5
Rule 5: Links between people unrelated to the head						
	Any non-relative age 0-15	Unmarried partner	15-44	strict	yes	5
	Any non-relative age 0-15	All other non-relatives	15-44	strict	yes	5

1. When the household head is polygamous, we narrow the allowable age difference between a potential mother and child and give priority to women who have not exceeded their child cap. Children who do not link to any mother, are linked to the head and to his first spouse.

2. When two children/children-in-law are linked by SPLOC, the first listed receives parent link

3. Applies only to France 1962-1975

4. In samples with childbearing data, a potential mother must be ever-married or in a consensual union or have ever given birth. A potential father must be ever-married or in a consensual union.

5. Number of links cannot exceed a woman's observed number of children-ever-born or constructed child cap. A potential father must be ever-married or in a consensual union, and number of links is limited based on spouse/partner's childbearing history.

6. Allowed only in samples without a parent relationship code

7. Allowed only in samples without a grandchild relationship code