Building Relationships Where There Are None: Imputing Relationship Status in the 1850, 1860 and 1870 Decennial Census Files

Josiah Grover, Jose Pacas and Steven Ruggles
Minnesota Population Center, University of Minnesota

Introduction

The U.S. 1850, 1860, and 1870 Decennial Censuses did not include a question about relationship to head of household which is a major limitation for the study of historical family structures. The IPUMS project overcomes this lack of data by exploiting the 1880 census microdata as a donor pool for relationship status in these earlier years, making IPUMS USA the only source of family interrelationship data for the time period. Although the imputed data have been available since 1995, there has been little documentation on the procedure. Moreover, since 1995 there have been improvements to the process for imputing relationship status as well as technological innovations that have changed the resulting imputed values.

In this paper we have three main goals. First, we outline the methods used to impute relationship status for the microdata samples from 1850 through 1870. These include five 1% samples, two of which include oversamples of the African American population, as well as one complete count microdata file with the two remaining complete count files expected in the near future. We then highlight specific complications of working with these early datasets. Lastly, we document the reliability of these imputed relationships and introduce new ways of testing their reliability. We test our procedure by comparing imputed relationships to true relationships in the 1880 microdata sample. We plan to extend this analysis to the 1900 and 1910 Census microdata samples, both of which have collected data on relationship status directly. By including these extra years of data, we are able to add a temporal component to our validation which allows us to analyze trends as well as cross-sectional distributions and mismatch rates. We can also estimate how much the mismatch rate grows as we move further from the 1880 source of the donated values.

In sum, our paper documents and tests the longstanding practice of imputing relationship status in the 1850 through 1870 datasets.

Background

Only limited documentation has been available for IPUMS family relationship imputation. Ruggles (1995) briefly described the procedures for imputing family relationships in the 1850, 1860 and 1870 samples.[1] Additional detail was published as part of the core IPUMS documentation in 1998, and has been available since then on the IPUMS website. In our paper, we fully document improved procedures used in the newest IPUMS data releases. In addition, we provide detailed analyses of how well our imputations match reported relationships. The following section will set the stage for the sort of analysis we undertake.

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

For the purposes of this abstract, we give a general overview of the imputation procedure. The primary goal is to assign each person one of the following relationship-to-householder values (IPUMS variable: RELATE).

Table 1 - Relationship Categories

Relationship to Household Head Categories	
RELATIVES	
01	Head/Householder
02	Spouse
03	Child
04	Child-in-law
05	Parent
06	Parent-in-Law
07	Sibling
08	Sibling-in-Law
09	Grandchild
10	Other relatives
NON-RELATIVES	
11	Partner, friend, visitor
12	Other non-relatives
13	Institutional inmates

IPUMS does make more detailed relationship values available but not in the earlier samples. For the purposes of imputation, we do not assign anyone the “Partner, friend, visitor” but rather collapse these into the “Other non-relatives” category.

Our process is separated into two steps: Logical rules and probabilistic imputations. The logical rules exploit the 1850-1870 census instructions which specified to enumerators that: "the names are to be written beginning with the father and mother; or, if either, or both, be dead, begin with some other ostensible head of the family; to be followed, as far as practicable, with the name of the oldest child residing at home, then the next oldest, and so on to the youngest, then the other inmates, lodgers and boarders, laborers, domestics, and servants" (https://usa.ipums.org/usa/chapter5/chapter5.shtml). The logical rules identify householders, spouses, and children, accounting for about 75 percent of the cases in 1880, with an overall error rate (in 1880) of under 1%.

An actual example of these sorts of households are the Smiths. We see a married couple, Neil and Kitty, 24 and 21 respectively, with three children. The household contains no other families or people, all share the same surname and the children are listed from eldest to youngest.

Table 2 - One of the Smith Families in 1880

Last Name	First Name	Age	Relate
SMITH	NEIL	24	1. Head
SMITH	KITTY	21	2. Spouse
SMITH	HOLMAN	7	3. Child
SMITH	IOLA	6	3. Child
SMITH	CORNELIA	1	3. Child

When an individual cannot be assigned through logical rules (about 25% of cases), we allocate their relationship through imputation. For these, we designed a probabilistic "hot deck" imputation procedure similar to the procedures that the Census Bureau uses to allocate missing and inconsistent information. We use nineteen key individual characteristics available in the 1850-1870 samples that were strong predictors of family relationship in 1880. One example of an imputation is Charlotte Holifield. She is listed at the end of a household which contains the Madden family. Notice that the Madden family was assigned using logical rules. There is a head, a spouse, and children who are listed eldest to youngest and all share a common surname. So who is Charlotte Holifield? Based on our imputation, we assign her a relationship of Other Non-related individual. This case highlights how our predictors use variable such as occupation to assign relationship. Here we can see that Charlotte was listed at the end of the household roster as a domestic servant. Our predictor score would match her to other domestic servants and thus assign her a relationship of Other Non-related individual.

Table 3 - The Madden Family in 1880

Last Name	First Name	Age	Occupation	Relate
MADDEN	JESSIE C.	42	Retired	1. Head
MADDEN	SIDIN A	39	Keeping house	2. Spouse
MADDEN	GEORGIA A.	15	Student	3. Child
MADDEN	JOSEPH E.	13	Student	3. Child
MADDEN	MATTIE B	9	Student	3. Child
MADDEN	JESSIE R.	7	Student	3. Child
MADDEN	RICHD. A	4	Blank	3. Child
MADDEN	S. VIRGINILA	2	Blank	3. Child
HOLIFIELD	CHARLOTTE	19	Domestic Servant	12. Other non

In our paper, we will give a detailed overview of the procedure used for imputing relationship to household head.

Assessing the Reliability of Imputed Relationship

At the core of our paper is testing how well we do at imputing relationship. Here we preview the sort of analysis we will undertake. In general, we are looking to maximize the match rate between a person's reported relationship and their imputed relationship while also matching the overall distribution of relationship in the sample. These two goals can at times be at odds with each other. Improving the match rate on less common relationships (like siblings) may result in a less accurate distribution. We will elaborate on this dynamic in the paper. For this abstract, we compare the imputed and actual relationships in the 1880 1% file.

Individual-level match rates

The first measure we look at is the match rate between individual responses and imputations. As seen in Table 4 below, the overall match is about 94%. There is considerable difference between the match rates for those assigned using logical rules and those that are imputed. Indeed, our logical rules are over 99% accurate and accounts for over three-quarters of our sample. Our match rate is diminished for the imputations at around 72% but accounts for a substantially smaller proportion of our overall sample. These results reflect a common theme in our research: We are able to almost perfectly assign simple relationships (head, spouse, child) but are less accurate as relationships become more complicated.

Table 4 - Match Rate by Relate (12 Categories)

Imputation Method	Match Rate	Percent of Sample	Count
All	93.68%	100%	502,819
Logical imputation rule	99.14%	77.6%	390,112
Hot deck imputation	72.45%	22.4%	112,707

These match rates use the RELATE described above with twelve categories. Using a simpler relationship variable of just five categories (Head, spouse, child, other relatives, and other non-relatives), we have an even higher match rate.

Table 5 - Match Rate by Broad Relate (5 Categories)

Imputation Method	Match Rate	Percent of Sample	Count
All	94.66%	100%	502,819
Logical imputation rule	99.49%	77.6%	390,112
Hot deck imputation	77.96%	22.4%	112,707

We also break down our match rates by the broader relationship categories. These rates confirm that our imputations are extremely accurate at capturing Heads, spouses and children but much less so with the other relationships which require hot deck imputation.

Table 6 - Match Rate within Broad Relate Categories

Relate	Match Rate	Count
Head	**99.97%**	**101,605**
Spouse	**98.81%**	**80,627**
Child	**97.18%**	**239,420**
Other rel.	**64.74%**	**19,874**
Non Rel.	**81.05%**	**34,459**

Distributional similarity

Next we compare the distributions based on the detailed relationship categories as well as the broader categories. As Figure 1 shows, the imputed relationship results in a distribution that is nearly identical to that of the actual relationships. In comparing how much each relationship category deviates from what the distribution should be, we see that our largest source of error comes from the Parent-in-Law category, followed by the Sibling-in-Law and Siblings, and then by Child-in-Law. This pattern emphasizes the difficulty in correctly imputing relationships where the surname tends not to match that of the head of the household.

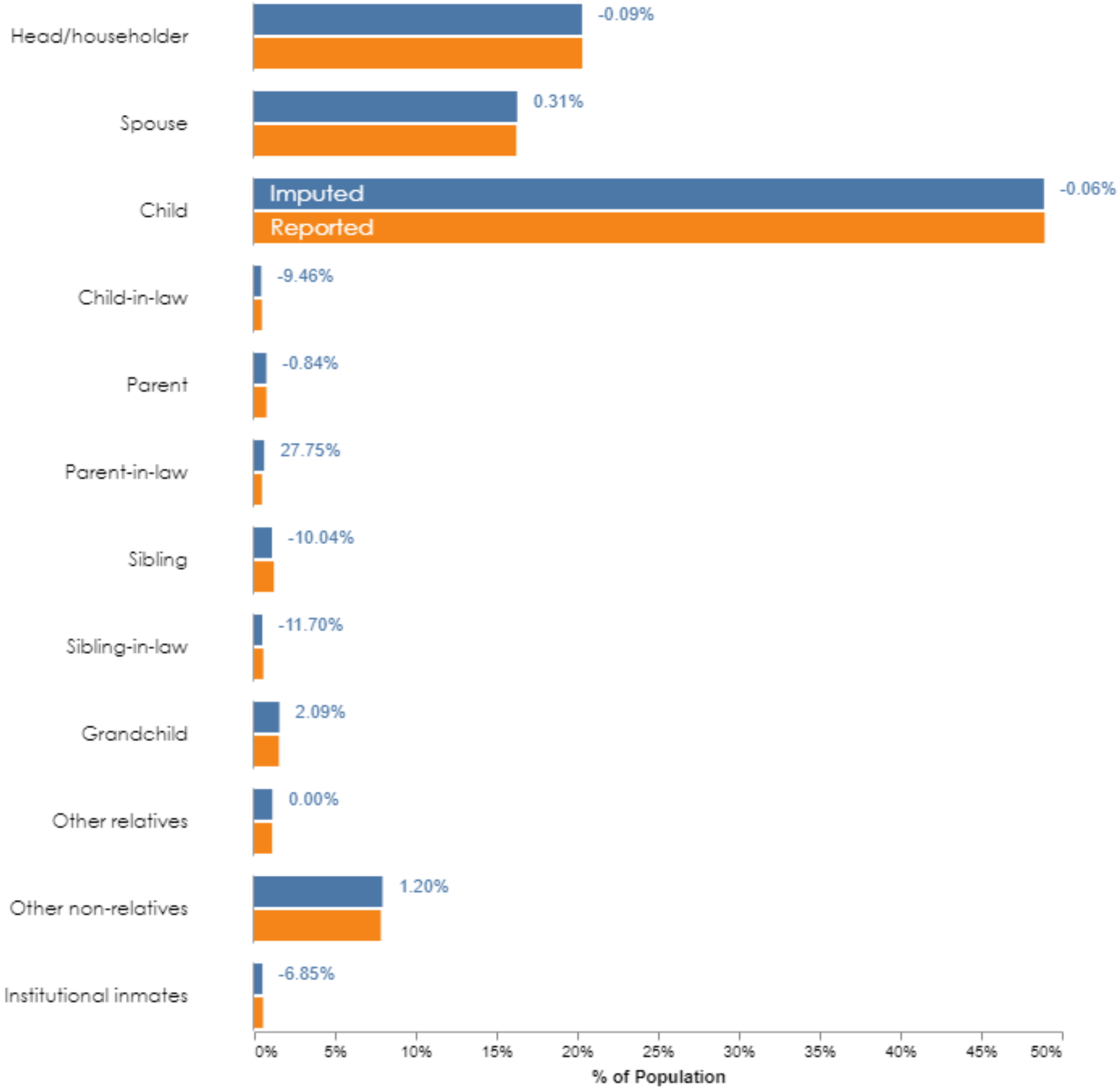

Figure 1 - Distribution of Relate - Imputed v. Reported

If we broaden the relationships into five categories, our distributions are much closer. As Figure 2 shows that each relationship category differs at most by 1% from the actual relationship. In sum, we are able to closely match the distribution of relationships. The relationship types that we fail to closely match are relatively smaller populations and thus tend to not throw off the overall match rate very much. In sum, our paper will take an in-depth look at how well we impute relationship in the 1880 1% sample using the sort of metrics presented here.

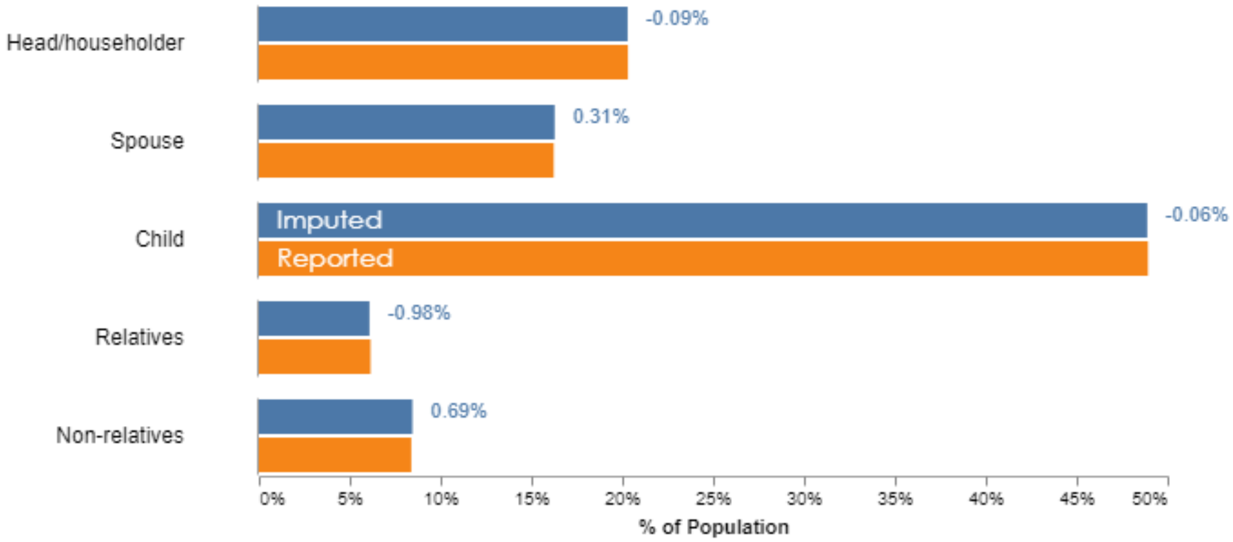

Figure 2 - Distribution of Broad Relate - Imputed v. Reported

Relationship Distributions Across Time

The goal of these analyses is to measure how well our imputation works within a sample that has actual relationship data but to the end of applying our imputations back to 1850, 1860, and 1870. We expect that our accuracy will diminish the further we get from our focal sample of 1880. In order to get an idea of the magnitude of this inaccuracy, we extend our imputations to include 1900 and 1910, two samples which also collected actual relationship information. Note that there exist no 1890 data. Using the broader five category relationship, we plot out our distribution from 1850 through 1910 in Figure 3. Focusing first on the deviations from the true distribution, we see that our imputation tends to impute less householders and children and imputes more spouses overall as we move towards 1910. These deviations are not large, however. The Other Relative and Non-Relatives categories have the highest deviations by far. We tend to impute too few Other Relatives and too many Non-Relatives. Overall, these findings are very promising in that we do not identify large deviations.

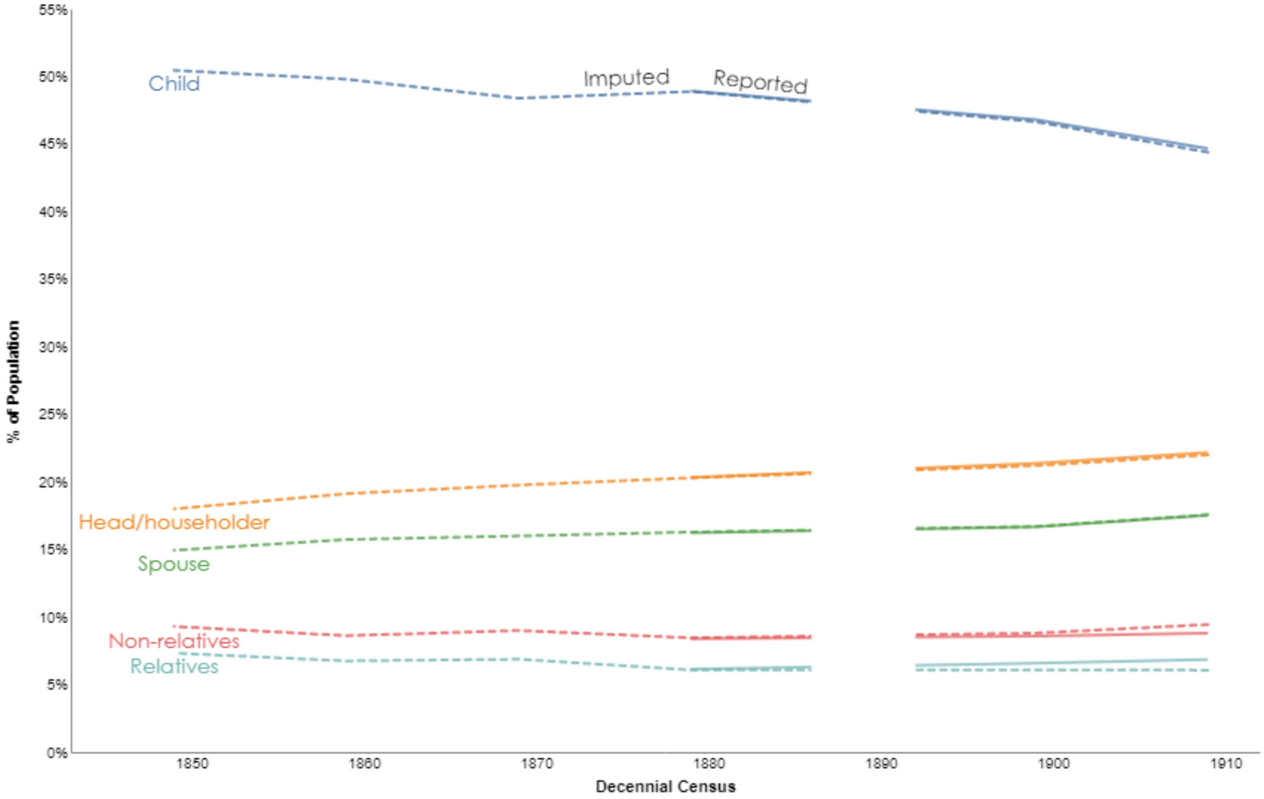

Figure 3 - "Broad" Relate as percent of total population - Imputed v. Reported - 1850 – 1910. (Note: The 1890 Census files are not available.)

Finally, we use one last metric to look at patterns over time. Figure 4 (next page) plots the ratio of Related Individuals to Non-Relatives. We see a large divergence after 1880 between the reported relationships and our imputed relationships. By this metric, our imputations seem to be getting worse over time. With this sort of information, we can look to improve our procedure to more closely match the reported ratios and/or provide the researcher with an idea of the errors over time. Ultimately, we aim for transparency and providing as much information as possible.

Future Steps

We will extend our paper in three main ways. First, we will expand on the procedure of the imputation with much more details. Second, we will expand our analysis on the reliability of our imputations. In particular, we will identify particular relationships where the imputations can be improved. Third, we will explore whether we can improve the donor pool used for the hot deck imputation. Currently, this donor pool consists of about 112 thousand people who are other relatives and non-relatives from the 1880 1% sample. We plan on creating a donor set from the 1880 5% sample and testing the reliability of our results with this larger donor set.

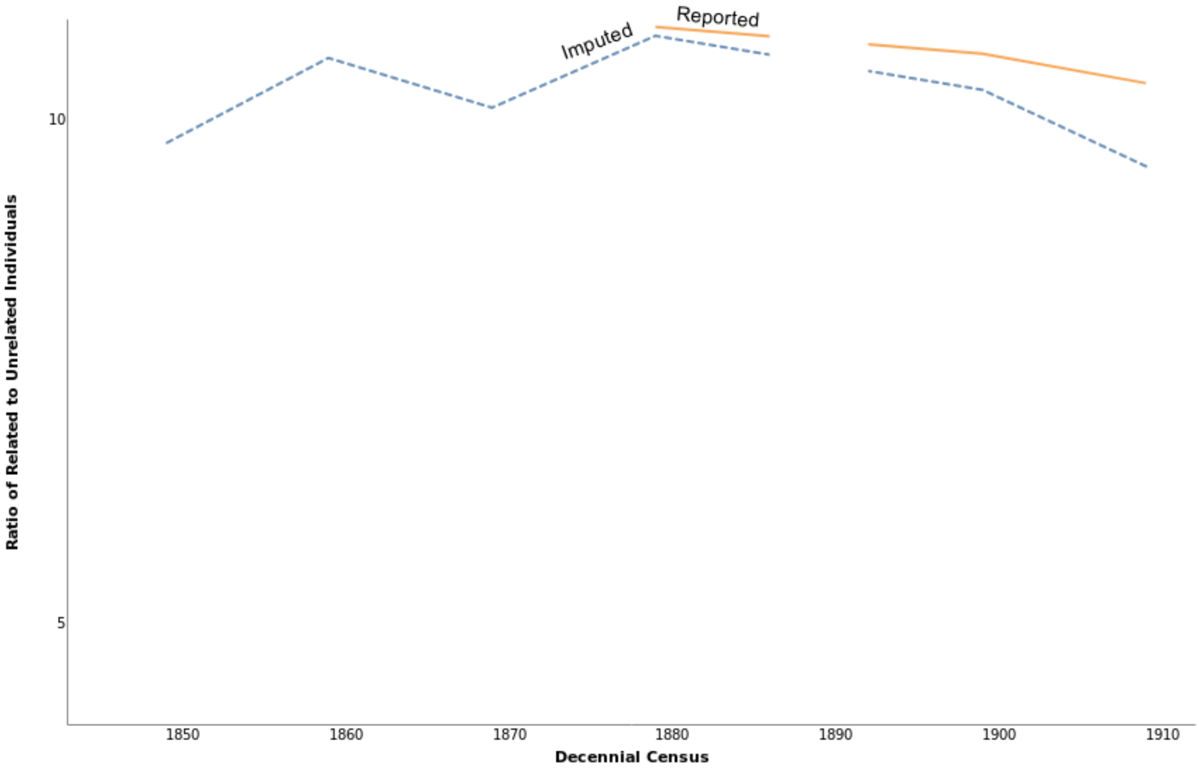

Figure 4 - Ratio of Related to Unrelated People - 1850-1910