The use of linked census data to estimate mortality and fertility in the pre-registration era of the United States: An application of the 1850-1940 IPUMS Multigeneration Longitudinal Datasets

Jonas Helgertz and J. David Hacker, Lund University and University of Minnesota

Introduction

In this paper we describe a method to estimate child mortality with panel datasets of individuals linked between the 1850-1880 and 1900-1940 censuses of the United States. The large number of cases in the datasets allows us to estimate child mortality in all states and in most counties for the white and black populations for each of the seven decades covered by the datasets. The method also allows us to estimate of census undercounts among children. With these estimates we make two additional contributions. First, with the assistance of a Brass model life table system, we construct new life tables indexed to the child mortality estimates by county and decade. We show that geographic variations in mortality were large (hardly a surprise to historical demographers working in other contexts, but useful for U.S. historical demographers limited to national life tables). Second, with the help of these new life tables, we construct new estimates of fertility, also at the county level, using the IPUMS 1850-1880 and 1900-1940 complete-count datasets and Own Child Methods (OCM) of fertility estimation. We estimate general fertility (age-specific birth and total fertility rates), marital fertility (age-specific marital fertility and total marital fertility rates), and Coale and Trussell indices of fertility, marriage, and marital fertility (I_f, I_m, I_g, M, and m).

These estimates, we believe, represent a major increase in our understanding of the timing and spatial variations in the demographic transition in the United States and will prove to be a valuable resource for other scholars (we intend to share our county-level estimates online through the IPUMS MLP website for interested researchers). We also believe the method to estimate child mortality will be useful for mortality analyses at the individual level, so we describe them in enough detail to allow replication by other researchers. Our major goals in this paper are to describe our methods and present an overview of our results. Although we do not show all these mortality, fertility, and undercount estimates—which together comprise a large dataset—we map selected child mortality, fertility, and census undercount estimates by county for a few census years and discuss the major implications of our descriptive results.

Background

The United States was late to establish a national system of vital registration. Although a few states and municipalities began registering births and deaths in the mid nineteenth century, the U.S. Death Registration Area (DRA) was established only in 1900 and initially included just 10 states and the District of Columbia, together representing just 26.3% of the nation's population. At its founding, the DRA was unrepresentative, being comprised predominately of states in the Northeast census region that were more urban and industrial than non-DRA states and with higher proportions of foreign-born residents and lower proportions of Black residents. States were added to the DRA over time, making it more representative of the nation by the 1920s, but it was not completed until 1933, when Texas was added. Even then a significant percentage of deaths are believed to have been unregistered. The national Birth Registration Area began even later, in 1915, and births continued to be underreported after it was deemed complete in 1933 (Preston and Haines 1991: 49-50; Haines 2000: 311-313).

Because of the lateness and deficiencies of vital registration, our understanding of the mortality and fertility transitions in the United States, especially in their early stages, rests on weak empirical foundations. Ironically, about the time the U.S. vital registration system was completed in the mid-1930s, the nation's century-long fertility transition came to an end with the commencement of the Baby Boom. Although the mortality transition continued, the years of the greatest reduction were in the half century prior to 1933. Researchers of the demographic transition have needed to be creative, constructing demographic rates using census data and indirect estimation methods or from studies of genealogical records, the available vital registration data, and special data sources, including retrospective mortality information collected with the 1850-1900 censuses and children ever born and children surviving information collected for each mother in the 1900 and 1910 censuses (Preston and Haines 1991).

The most commonly cited life tables for the nineteenth century, Michael R. Haines' (1998) national life tables for census years 1850, 1860, 1870, 1880, 1890, and 1900, were based on retrospective mortality censuses taken in those years (households enumerated in the population census were to report any household members who had died in the year prior to the census). These censuses clearly under-reported deaths—perhaps as much as 40% or more overall—but Haines relied only on children aged 5-14 dying in the year prior to each census, who appeared to be well reported. He then used model life tables to extend the estimates to all age groups. Hacker (2010) also used model life tables to construct nineteenth-century life tables, fitting adult life expectancy estimates at age 20 from published genealogical studies to a relational life table system based on the 1901 DRA. No attempt was made by Haines or Hacker to estimate regional or subregional estimates mortality or life expectancy. Both sets of life tables were also for the white population, although Haines included a few life tables for the Black population after 1900 based on his analysis with Samuel Preston of the children ever born and child surviving data in the 1900 and 1910 censuses.

Although the U.S. lacked a system of vital registration until the early twentieth century, it conducted decennial censuses beginning in the late eighteenth century, which can be used to make indirect estimates of mortality and fertility. The population of the Black population was essentially closed to migration after the abolition of slave imports from the trans-Atlantic slave trade in 1808, allowing mortality to be estimated with two census methods and assumed model life table (Eblen 1974). Differences in census coverage errors, age misreporting, and the unknown numbers of slave smuggled into the nation after the cessation of the international slave trade, however, limit confidence in two-census mortality estimates (McDaniel and Gruska 1995). The white population, of course, was continually augmented by new immigrants and not closed to migration. Two census methods are only feasible for the native-born white population (which can be assumed closed to migration) after 1880, when the Census Office cross-tabulated the population by age, sex and nativity.

With an assumed age structure of dying, intercensal growth rates, and stable population methods, census data can be used to estimate crude birth and death rates (Eblen 1974, Yasuba 1962). Coale and Zelnik (1963) interpolated between the few available nineteenth-century life tables back project annual births from the age distributions of the native-born population in the 1880-1950 censuses. Aggregated counts of the population cross-tabulated by age group and sex can also be used to calculate child woman ratios, which are a good proxy for fertility. Between 1800 and 1860 the Census Office cross-tabulated the white population by age and sex for each county in the United States, facilitating cross-sectional studies (Easterlin et al. 1978; Hacker et al. 2021).

The construction of IPUMS microdata samples of the surviving 1850-1880 and 1900-1930 censuses in the 1990s provided new research opportunities. Although the Census Office did not cross-tabulate the white population by age, sex and nativity until 1880, researchers can use the 1850-1870 IPUMS samples to do so, extending two-census mortality estimates back to 1850 (Hacker 2010). More importantly, the microdata samples allowed the application of Own Child Fertility Methods (OCM) to estimate age-specific general and marital fertility rates. These estimates indicate that the marital fertility transition at the national level did not begin as early as the date implied by child-woman ratios (Hacker 2003). The decline began in New England in the mid nineteenth century and lagged in the South, which did not experience a significant decline in marital fertility until late in the century These OCM estimates, however, were based on the assumption that available national life tables (Haines 1998; Hacker 2010) accurately represented child mortality conditions in all areas of the nation. Hypothetically, if child mortality was higher in New England and lower in the South, regional differences in fertility would not have been so profound. Although the introduction of new complete-count IPUMS datasets (Ruggles et al. 2020) represents an ideal opportunity to apply OCM to smaller areas of geography, our lack of knowledge about local mortality conditions will limit confidence in these estimates. Ideally, therefore, we need a method to estimate mortality differences for small areas that can in turn be used as an input to OCM estimates. The IPUM MLP datasets provide that opportunity.

Data

We constructed panel datasets using preliminary "crosswalks" published by the IPUMS Longitudinal Multigenerational Project (IPUMS MLP) at the University of Minnesota (Ruggles et al. 2021). The crosswalks identify individuals linked between adjacent censuses in the IPUMS complete count datasets of the 1850-1880 and 1900-1940 decennial censuses (see Helgertz et al. 2020 for a description of the linking methods). We used the restricted versions of the complete-count datasets, which includes individuals' first and last names, and selected only married couples of childbearing ages linked between censuses. We also selected all children of the linked couples in both census years, whether the children were linked by the IPUMS MLP or not linked. We created three panel datasets for the nineteenth century (1850-1860, 1860-1870, and 1870-1880) and four panel datasets for the twentieth century (1900-1910, 1910-1920, 1920-1930, and 1930-1940).

A recent comparisons to couples in the linked dataset to couples in cross-sectional complete county datasets indicates that linked couples are fairly representative, with some modest bias in expected areas (e.g., couples appear to have been slightly more likely to be linked if they were literate, wealthy, native born, living in rural areas, living on a farm, living in the Northeast census region, and if they had more children living in the household) (Hacker et al. 2023). Overall, however, the biases appear small. Our selection criteria, however, was likely correlated with child mortality: because our method relies on linked couples, who (by definition) survived the decade under observation, children of linked couples likely experienced lower mortality than children of parents whose marriages were disrupted by the death of the father or mother. On the other hand, because our method assumes that children who could not be linked between censuses died in the interval, any missed links will bias our mortality results upwards.

Table 1 shows the number of couples in each panel dataset, the number of their children aged 0-3 and 0-9 in the first of the two linked censuses, the number of their own children aged 10-13 and 10-

19 in the second of the two censuses, and other characteristics of the datasets. [We will provide additional description of the dataset here, including Table 1]

Methods

The method is complex in detail but easy to describe in brief. Because the IPUMS MLP uses a high matching threshold to identify the same individuals in different censuses (with the objective of minimizing type I errors of declaring different individuals as the same individual), many linked couples have unlinked children in both censuses that would have been linked with a lower match threshold. Because we know these are children of the same married couples, lower matching thresholds can be used without incurring significant type I errors.[1] In a first step therefore, we use Jaro-Winkler distance scores of children's names, and their ages, sexes, birthplaces, and other information to make the best possible matches between the children of linked couples aged 0-9 in the first of the two linked censuses (hereafter "Censuses A") to the children of the same linked couple aged 10-19 in the second of the two linked censuses (hereafter "Census B"). For children in Census A who could not be linked to Census B, we expanded our search to include individuals outside of linked couples' households with similar years of birth, names, and birthplaces, but reverted to higher matching thresholds. After identifying all linked children, we next identified deceased children in Census A and under-counted children in Census B. We assumed children aged 0-9 in Census A who could not be linked forward to an individual aged 10-19 in Census B died during the ten- year interval, while all children aged 10-19 in Census B who were not linked backward to a child aged 0-9 in Census A were undercounted by Census A. After analyzing and comparing the results to model life tables, which suggested an increasing tendency for children to leave their parental homes after age 13, we limited our analysis to children aged 0-3 in Census A.

We implement these steps in STATA. In the following section, we provide more detail/describe the code. We organize this discussion under three separate headings: (1) estimating 10-year forward survival ratios among children aged 0-3; (2) estimating census undercounts among children aged 0-9; (3) fitting model life tables for the white and black populations; (4) fertility estimation using own-child fertility methods,

Estimating 10-year forward survival ratios among children aged 0-3

[Detailed description here]

[1] Our 1870-1880 panel dataset, for example, included the 1870 Dent County, Missouri couple Joab and Elizabeth Hobson, aged 43 and 38 respectively, who were linked to the 1880 Benton County, Missouri couple Joab and Elisabeth Hobson, aged 53 and 48. Joab's state of birth was listed as North Carolina and Elizabeth's as Kentucky in both censuses. Although residing in different Missouri counties, the MLP linking procedures determined that this was the same couple. In 1870, Joab and Elizabeth had three male children, named "A L," "A J," and "F S," aged 4, 4, and 1, all born in the state of Missouri, who were not linked forward to a child in the 1880 census. In 1880, Joab and Elizabeth had three male children, named "Abraham Hobson," "Andrew Hobson," and "Phillip Hobson," aged 14, 14, and 11, all born in Missouri, who were not linked backward to a child in the 1870 census. Although the IPUMS MLP project used conservative rules to minimize type I errors, and therefore avoided linking these children, it seems highly likely that these children were the same children in both censuses.

Estimating census undercounts among children aged 0-9

[Detailed description here]

Fitting model life tables for the white and black populations

[Detailed description here]

Fertility estimation using own-child fertility methods

[Detailed description here]

Results

[We will describe our results here. We now have fully working code and have produced all estimates (child mortality, full life tables, and fertility) for three of the seven decades. The estimates appear excellent, but we are in the produce of trouble shooting a few outliers. We anticipate have full results and a working paper in the next three months. In other words, we anticipate no difficulty finishing the paper well before the August 2023 conference. To support this proposal, we show preliminary maps of fertility in 1870. (Figure 1).