IPUMS
Working Papers

IPUMS Full Count Datasets of the U.S. Censuses of Population 1850-1880

Matt A. Nelson†
University of Minnesota

Diana L. Magnuson
University of Minnesota

J. David Hacker
University of Minnesota

Matthew Sobek
University of Minnesota

Lap Huynh
University of Minnesota

Evan Roberts
University of Minnesota

Steven Ruggles
University of Minnesota

June 2024

IPUMS Full Count Datasets of the U.S. Censuses of Population 1850-1880

Matt A. Nelson, *University of Minnesota* (0000-0002-8849-4628, nels5091@umn.edu)
Diana L. Magnuson, *University of Minnesota* (0000-8729-5109, magn0031@umn.edu)
J. David Hacker, *University of Minnesota* (0000-0002-5971-955X, hacke010@umn.edu)
Matthew Sobek, *University of Minnesota* (sobek@umn.edu)
Lap Huynh, *University of Minnesota* (huyn0064@umn.edu)
Evan Roberts *University of Minnesota* (0000-0001-5621-4823, eroberts@umn.edu)
Steven Ruggles *University of Minnesota* (0000-0001-5353-2578, ruggles@umn.edu)

Abstract:

IPUMS has finalized databases for each of the United States population censuses from 1850 to 1880. These data are the result of collaborations between FamilySearch and Ancestry.com, which provided the raw data, and IPUMS, which enhanced the data with editing, standardized coding, inter-census harmonization, and documentation. We discuss the data capture process conducted by the nineteenth-century United States Census Office, construction of the modern datasets, and variable availability. We conclude by briefly discussing the potential and limitations of these data for social science research. The public data are distributed by IPUMS and available for researchers to use free of charge.

Introduction

In October 2001, IPUMS released a preliminary population database for all individuals in the 1880 Census of the United States (Goeken et al. 2003). Containing 50 million person records, its sheer scale was revolutionary. Although limited in some respects—the database was constructed from data collected by volunteers at the Church of Latter Day Saints without IPUMS oversight and lacked a few important variables such as occupation, literacy, and school attendance—the IPUMS 1880 full count dataset was eagerly received by academic researchers, who were able to pursue new lines of social research. The 1880 full count dataset, for example, facilitated the construction of summary files for contextual and spatial analysis and the linking of individuals to sample data from other censuses (Goeken et al. 2011).

In the more than two decades since the release of the full count 1880 dataset, IPUMS has constructed nine additional full count databases for census years 1850-1950 (excluding 1890, the manuscripts for which were destroyed in a fire). This paper describes the recently completed IPUMS full count datasets for the 1850, 1860, and 1870 censuses and additions and enhancements to the 1880 full count dataset. We describe each dataset's method of construction and variable availability, and note their strengths and weaknesses relative to pre-existing IPUMS low-density samples. Because the content and accuracy of the 1850-1880 datasets reflects decisions made by the nineteenth century Census Office and field enumerators, we begin by reviewing the original data capture process, focusing on changes in the administration and enumeration of the census between 1850 and 1880.

All IPUMS full count population datasets have been constructed with considerable collaboration and investment from private genealogical companies. To protect their investments,

some information of interest to genealogical customers (primarily individuals' first and last names) are not included in the public-use versions of full count datasets distributed by IPUMS (http://ipums.org). Restricted-use versions containing all information can be obtained by researchers with approved research projects who agree to non-disclosure requirements. The IPUMS Multigenerational Longitudinal Project (IPUMS MLP), for example, relies on names—together with other identifying information such as age, race, sex, and birthplace—to link individuals across multiple censuses (Helgertz et al. 2022). Recent and future enhancements to the IPUMS MLP datasets and other linking projects requiring non-restricted use data are discussed in this special issue of *Historical Methods*.

Original Data Capture: The U.S. Census, 1850-1880

The U.S. censuses in the late nineteenth century must be understood in the context of the wrenching social, political, and economic upheaval surrounding the American Civil War and Reconstruction. The sectional crisis that dominated the national scene at the time of the 1850 and 1860 censuses is clearly evident in the *Congressional Record,* as efforts to reform the administration and content of the census were discussed amidst acrimonious debate between North and South over the expansion of slavery. In the aftermath of the American Civil War, further efforts to reform how the census was conducted, both in the field and administratively in Washington, played out amid agonizing post-war Reconstruction (Anderson 2015). The content and accuracy of the 1850-1880 datasets created by twentieth and twenty-first century social scientists reflects decisions made by the nineteenth century Census Office and field enumerators. We thus begin by reviewing the original data capture process, focusing on changes in the administration and enumeration of the census between 1850 and 1880.

The limitations of the Census Office to conduct an accurate count became more apparent with each succeeding census during the period 1850-1880: the protracted enumeration process (which increased the likelihood of missing or double-counting individuals); the lack of a permanent census office (which hindered the accumulation of institutional knowledge); supervision by U.S. Marshals who had competing duties; the formation of enumeration subdivisions that unevenly distributed the work; minimal training and oversight of assistant marshals; and the meager compensation of canvassers (which limited the attractiveness of the work to well qualified candidates). The Superintendent of the 1850 census complained that assistant marshals were selected not "for their especial fitness" for the job, but rather because they were "willing to undertake it." The Superintendent further complained that "An examination of the returns and the correspondence of the office" showed that "capacity" for census work among marshals and their assistants "was often the exception" (DeBow 1854). According to the November 1860 report of the Superintendent of the Census, the 1850 instructions were "not interpreted alike by all who executed them," resulting in "some want of harmony in the returns" (Kennedy 1860: 501).

The 1850 census marks a turning point in the history of the decennial census of population, not only in terms of the population schedule itself but also in the administrative organization of the census. The 1850 census was the first to have a formalized Census Board overseeing pre-enumeration issues and a hierarchical administration overseeing the enumeration process. Accordingly, the census was expanded to include six separate schedules–free population, slave inhabitants, mortality, agriculture, manufacturing, and social statistics–up from three schedules in 1840. As in the previous decennial censuses, many people successfully influenced the specific content of the enumeration schedules, including statistical experts, political leaders, and the

government officials tasked with carrying out the census (Magnuson 1995, 24-45; 9 Stat. 402 1849).

Statistical experts outside of government urged Congress to organize professional administration to oversee all census related work. What Congress assembled was much more modest; a Census Board charged with examining and reporting on various possible features (DeBow 1853, 13). The Census Board consisted of the Secretary of State, the Attorney General, the Postmaster General, and a secretary appointed by the Board (9 Stat. 402 1849; Anderson 2015, 43-44; Magnuson 1995, 33-35). What emerged from the Board's efforts was a radically different enumeration schedule that in turn greatly increased post-enumeration clerical work of the Washington Census Office staff (Anderson 2015, 42-52; Ruggles and Magnuson 2020). Meanwhile, the Senate created a select committee to draft a census bill. Unsurprisingly, miscommunication and lack of communication ensued, and after "torturous" debate, the 1850 Census Act was passed just eight days before the enumeration was to begin (Anderson 2015). If no other provisions were made for the 1860 or subsequent censuses, later censuses were also to be taken under the Census Act of 1850 (9 Stat. 428 1850).

The 1850 Census Act transformed the administration of the census as well as its form and content. Oversight of all stages of the census was moved to the Secretary of the Interior (9 Stat. 395 1849). Most significantly, the 1850 Census Act shifted the unit of enumeration from the household to individual-level enumeration. This was a monumental transformation that later statisticians would deem "an epoch in the history of census taking in this country" (Wright and Hunt 1900: 47). The 1850 form required the assistant marshal to list each individual on a separate line and fill in answers for fourteen questions per person, rather than capturing information at the

household level, as in previous censuses. Thus the potential for future individual-level microdata was born.

Refinement to enumeration procedures in 1850 reflected expansion of the decennial census and population growth. Instructions outlining duties and responsibilities of the U.S. Marshals and their assistants were included in the 1850 Census Act, and for the first time specific procedural enumeration instructions for the canvassers were distributed by the Census Office (Wright and Hunt 1900: 148-150). Assistant marshals received an example copy of a correctly returned enumeration form and could expect that their respective U.S. Marshal would check in and review the progress of their work. U.S. Marshals were directed to divide districts into subdivisions not exceeding 20,000 people. Completed schedules were to be carefully examined by the U.S. Marshals and returned for correction to the appropriate assistant marshal if discrepancies were detected. Assistant marshals were required to submit the original enumeration schedules to their respective county clerk and send two copies to the U.S. Marshal of their district (9 Stat. 428 1850). Despite efforts to provide training and oversight for the U.S. Marshals and assistant marshals, the enumerators did not uniformly interpret the instructions. The superintendent of the census complained that the incompetence of enumerators could be blamed on the low rate of compensation, the appointment process, and the size of enumeration districts (DeBow 1853: iv).

The 1860 decennial census was taken under the Census Act of 1850. Instructions to the U.S. Marshals and assistant marshals were refined and elaborated upon to "overcome the difficulties which arose in the course of the enumeration in 1850 and to avoid all misapprehension as to the intent of the inquiries" (Wright and Hunt 1900: 51; U.S. Census Office 1860). When the final 1860 census volume was presented to Congress in 1864, the assessment noted the improvement over prior censuses but acknowledged the persistence of errors: "Every effort has

been made to insure accuracy, and, it is believed, with success. While errors may occur, it is confidently believed that they will be of minor importance, and less in number than have appeared in any previous census" (Kennedy 1864: iii). The 1860 census was completed before the outbreak of the American Civil War, and no schedules were lost or delayed in their return to Washington.

The 1870 decennial census was also taken under the Census Act of 1850 (minus the slave schedule), but not without a tremendous effort to revise the legislation before the enumeration was to commence (Magnuson 1995: 46-50; Anderson 2015: 80-81). A Special Committee on the Ninth Census was formed by the House of Representatives. Despite the efforts of the Committee, the outcome of the proposed 1870 legislation was far less than the Special Committee members and the outside statistical community had hoped. In addition to revising the population schedule, the Special Committee proposed to overhaul the census by addressing issues related to the employment of assistant marshals as enumerators, unequal enumeration district sizes, the protracted enumeration period, lack of confidentiality, and compensation for assistant marshals (H.R. No.3 1870). The lengthy report of the Special Committee and the backing of the national and international statistical communities could not overcome Congressional resistance and personal animosities. (Magnuson 1995: 111). With minor amendments, the census of 1870 was thus taken under the 1850 census act (Wright and Hunt 1900: 920).

Despite the failure of the new census legislation, the Ninth Superintendent of the Census, Francis A. Walker, initiated what he considered necessary reforms for the training and oversight of the U.S. Marshals and assistant marshals to improve the accuracy of the population count. Walker had been part of the Special Committee on the Ninth Census. In the course of serving twice as Superintendent, Walker developed significant statistical, economic, and public policy expertise (Anderson 2015: 81-82). Two important refinements were imposed by Walker in 1870 in an effort

to improve canvasser oversight and training: assistant marshals were directed to report their progress every fortnight to their respective U.S. Marshal; and canvasser instructions were made more precise than in previous censuses (U.S. Census Office 1870). The issue of a protracted enumeration was minimally addressed through an amendment to the 1850 Census Act, which required the population schedule to be submitted to the Census Office by September 10, 1870, just over three months after the June 1, 1870 census day (16 Stat. 118 1870). Despite the directive to compress the enumeration period so as to more accurately capture the American population as it was on June 1, 1870, the field work was not completed until August 23, 1871 (Wright and Hunt 1900: 55-56).

Oversight of the 1870 census provided Superintendent of the Census Walker with ample evidence to build his case for implementing the failed recommendations of the Special Committee at the next (1880) decennial census. Walker believed that the outdated census administrative machinery was to blame for the faults of the 1870 census, and his summary report made plain what he considered the necessary changes that needed to be made before the 1880 count: a shorter enumeration period to increase accuracy; replacing U.S. Marshals with canvassers under control of the Census Office; decreasing the size of enumeration districts; and raising the meager compensation of enumerators (Walker 1882: xxi-xxvi).

By 1879, two factors favored those advocating procedural reform of the decennial census of population. First, in the decade between 1869 and 1879, Francis Walker gained experience and credibility as the country's leading demographic expert (Munroe 1923). In addition, congressional sentiment toward procedural reform had caught up with the enthusiasm and concern of Walker and those represented by the Special Committee. When Senator James Garfield articulated his procedural reform proposal from the decade preceding (he had served on the Special Committee

for the 1870 Census), Congress supported his thinking and agreed that the census machinery needed revolutionary change (Magnuson and King 1995: 28).

The procedural changes enacted through the Tenth Census Act were indeed revolutionary. The 1880 census marked a watershed in local administration, enumerator selection, and oversight by the Census Office. The 1879 Census Act addressed the main concerns Francis Walker and the Special Committee had with the old 1850 Census Act. Congress replaced the U.S. Marshals with officials called "supervisors." The Secretary of the Interior was responsible for designating 150 supervisors, representing the states and territories, to be approved by the President. The issue of the protracted enumeration period was addressed by condensing the enumeration period to one month (June 1-July 1, 1880); two weeks were allowed for cities with over ten thousand inhabitants. Compensation of supervisors and enumerators was also increased, though the pay for the labor-intensive work was still considered insufficient. Finally, the Superintendent of the Census took charge of forming the supervisor and enumeration districts (20 Stat. 473 1879; Magnuson and King 1995; King and Magnuson 1993).

Placing census supervisors within the Census Office, itself housed within the Department of the Interior, tightened organizational control of their work in three respects. First, the authority to appoint and remove supervisors was "absolutely essential to good administration" (Walker 1883: xxxiv). Supervisors could now be expected to focus solely on census-related work for the duration of the count. One of the most important tasks of the census supervisors was the selection of enumerators. The census law and Census Superintendent set general guidelines for supervisors to follow when selecting enumerators. Implementation of this directive was uneven, but supervisors understood the bottom line: hiring an incompetent enumerator could negatively impact the community's count (Magnuson and King 1995; Magnuson 1995: 118-24, 136-38).

The second important effect of locating census supervisors and enumerators within the Census Office was the opportunity for greater oversight (Magnuson and King 1995). The appointment of supervisors directly answerable to the Census Office was an important structural change that implemented intentional oversight of fieldwork, backed by formal consequences for noncompliance (20 Stat. 473 1879). The Census Office and the district supervisors shared the responsibilities of training and overseeing enumerators, but it was the Census Office that developed the materials and the training plan. Once the enumeration commenced, a steady stream of communication between enumerators, their supervisors, and the Census Office ensued through completion of the count.

The third mechanism implemented by the 1879 Census Act to tighten oversight of enumerators and improve accuracy in the count was the use of smaller and more clearly defined enumeration districts, which doubtless reduced undercounts of the population. Prior to 1880, the Superintendent of the Census did not have the power to challenge the U.S. marshals' plans for dividing up the territory to be enumerated (Walker 1878). Enumeration areas were limited to 20,000 people in the 1850 Census Act. The 1879 Census Act cut the size down to 4,000 and Superintendent Walker further restricted the size of enumeration districts to no more than 2,500 residents. In practice, enumeration districts for the 1880 census were even smaller, encompassing 1,600 inhabitants on average. The task of defining enumeration districts fell upon census supervisors, but the basic rules for determining the boundaries were set by the Superintendent of the Census. Enumeration districts were to maintain the territorial integrity of counties and civil divisions. A civil division was to stand as an enumeration subdistrict if the population exceeded seven or eight hundred, but it could be divided to keep the population from exceeding 2,000 to

2,500. Clear boundaries like streams and roads were to mark off these divisions. And if a civil division were broken up, its subdivisions were to be of approximately equal population size.

Few records from the Census Office have survived, making it impossible to reconstruct all the directions coming from Washington on the training and oversight of enumerators. What is clear from newspaper reports across the country, however, was Superintendent Walker's determination to change the situation from 1870, when enumerators "were entirely independent" of his supervision and he "had no control of their work" (Magnuson 1995: 170-171). During the census period, supervisors were expected to be in constant contact with their enumerators, communicating "the necessary instructions and directions relating to their duties, and to the methods of conducting the census," as well as advising and counseling enumerators "in person and by letter as freely and fully as may be required to secure the purposes" of the law. The Census Act further directed supervisors to "examine and scrutinize" enumerator returns (20 Stat. 473 1879). Press reports indicate that the details of oversight varied across localities, ranging from direct supervisor review of schedules (Baltimore, Atlanta, and New York City) to large general meetings of enumerators (Philadelphia) (Magnuson 1995: 171-174).

A decade after the 1880 census, Walker summed up his view of the enormity of the task.

> The labor of organizing and energizing a census is such that no man can conceive who has not himself undertaken it, or, at least, stood close by and watched the machine in full operation. Aside from the question of the superintendent's intellectual ability to comprehend his work in all parts, and to make provision for every foreseen occasion and for every sudden exigency of the enumeration, the strain upon the nerves and the vital force of whomsoever is in charge of the census is something appalling. My successor in the Tenth Census, Col. Charles W. Seaton, was literally killed by the work, and three successive chief clerks of that census died in office … Superintendent of the Eleventh Census, Mr. Porter, was driven away to Europe by his physician … while the work was at its height, to save his life. Taking a Census of the United States under the present system, and upon the existing scale, is like fighting a battle every day of the week and every week for several months (Munroe 1923: 197-198).

Inevitably, decisions made by nineteenth century respondents, fieldstaff, and the Census Office all impact microdata quality of the IPUMS Full Count datasets. Nineteenth century administrative concerns about the length of the enumeration period, the lack of a permanent census office, unevenly sized enumeration subdivisions, and the use of temporary, minimally trained, and meagerly paid field staff all addressed factors that ultimately influence microdata quality. Modern day researchers cognisant of this historical context are best positioned to use IPUMS Full Count data to answer their research questions.

From Manuscript Returns to Social Science Data

The original census manuscript returns were preserved in the Commerce Building and later in the National Archives (Dorman 2008; Blake 1996). Use of the manuscripts by Census Bureau staff to support their Age Search service (established to assist citizens applying for Social Security to verify birth and citizenship status), meant the bound bulky enumeration forms were deteriorating from repeated use, especially the 1880 census. The unwieldiness of the bound volumes and sheer size of the decennial collection forced the Census Bureau to search for a long-term preservation solution. As a first step, under a Works Progress Administration funded project, Bureau staff transferred respondent information (name, age, sex, enumeration district and page number) onto index cards to increase efficiency for their Age Search operations. Next, in collaboration with two manufacturers of photography equipment, the Bureau developed microfilm cameras to handle different sizes and types of documents. Between 1937 and 1944 the Census Bureau undertook the enormous task of microfilming the manuscript census forms for 1840-1880 and 1900-1940 (Genadek and Alexander 2022; Dreiser 1944). In some cases, faded ink and difficulties focusing the camera on curved pages resulted in poor image quality (U.S. Census Bureau 2002: 111).

Historians soon discovered the research potential of the manuscript returns, which included some information that the Census Office did not tabulate for publication (e.g., the 1880 census collected information on marital status, but the Census Office did not tabulate the results). In a 1948 article in *Southwestern Historical Quarterly*, historian Barnes F. Lathrop called the unprinted census returns "a great neglected source," highlighting their potential for research on individuals and groups, small area studies, linking projects, and correlation analyses not attempted in the printed reports (Lathrop 1948: 293). Following Merle Curti's groundbreaking quantitative study based on manuscript returns of the 1850-1880 censuses in Trempealeau County, Wisconsin (1959), the "new" social and economic historians of the 1960s and 1970s turned to manuscript returns of the 1850-1880 censuses to examine socioeconomic mobility (Thernstom 1964), geographic mobility (Thernstrom and Knights 1970), the economics of slavery and slave ownership (Fogel and Engerman 1974), and other topics.

Quantitative studies of the late nineteenth-century U.S. population were greatly assisted by the construction of public use microdata samples at the University of Minnesota in the 1990s, beginning with a one-percent density IPUMS sample of the 1880 census (Ruggles and Menard 1990) and followed by similar IPUMS samples of the 1850, 1860, and 1870 censuses (Hacker et al. 1999). Among other topics, historians have relied on the 1850-1880 IPUMS samples to examine migration (Hall and Ruggles 2004; Rosenbloom and Sundstrom 2004), fertility decline (Hacker 2003), marriage, intermarriage, and divorce (Hacker 2008; Ruggles 1997; Wildsmith et al. 2003), occupational structure and labor force participation (Sobek 2001), intergenerational coresidence and family structure (Ruggles 1994), and Civil War mortality (Hacker 2011). The creation of the first full count dataset of the 1880 census (Goeken et al. 2003) stimulated more research, including studies on topics that required full count data, such as record linkage (Abramitzky et al. 2021),

kinship networks (Hacker and Roberts 2017; Nelson 2020), intergenerational mobility (Long and Ferrie 2018), and racial residential segregation (Grigoryeva and Ruef. 2015; Karbeah and Hacker 2023, Logan and Parman 2017).

From Low-Density Samples to Full Count Population Datasets

In the two decades following the initial work on the 1880 full count dataset, IPUMS has collaborated with genealogical organizations to produce full count census microdata collections of the United States spanning the period from 1790 to 1940. In 2009, IPUMS began a collaboration with The Church of Jesus Christ of Latter-Day Saints and Church History Department (today known as FamilySearch), to construct a full count dataset of the 1850 census. A few years later, IPUMS reached an agreement with Ancestry.com to provide data processing, cleaning, and coding in exchange for their full count transcriptions of the household-level censuses of 1790-1840 and individual-level censuses for 1850-1880 and 1900-1940. Data entry had commenced before the agreement by vendors under contract with Ancestry.com—primarily Beijing Formax—and was initially limited to variables most useful for genealogical research (e.g. name, age, sex, race, and birthplace). In 2014, the University of Minnesota signed another agreement with Ancestry.com to add most of the remaining variables for census years 1860-1880, the 1860 census of slave inhabitants, and the 1860-1880 censuses of mortality in exchange for monetary contributions to offset the costs of data entry. The added variables included schooling, literacy, occupation (1860-1870), wealth (1850-1870), disability and sickness (1880), citizenship, and denial of voting rights (1870), among others (Ruggles 2023).

The IPUMS collaboration with private genealogy companies resulted in considerable cost savings. In contrast to the original IPUMS samples, however, we had little control or oversight of the data entry process; nor did we have the resources to investigate many problematic cases or

manually code all unique occupation or birthplace strings. Inevitably, the automated methods required to clean and code these large databases resulted in some data quality compromises relative to the more handcrafted samples of the 1990s.

Data Processing

The data IPUMS received from Ancestry.com were literal transcriptions of the contents of the census forms, with most fields being character strings. In some cases, the transcriptions were standardized but generally were not. The data needed to be cleaned and coded to be useful for social science. The first step in processing the raw full count census data began with encoding geography and comparing aggregated case counts with population counts tabulated by the nineteenth-century Census Office. The level of detail in each published census volume expanded over the nineteenth century. For example, minor civil division population counts were not published in 1850 and 1860 but were included in the 1870 and 1880 census volumes. Published county population counts were available in digitized form, compiled by Michael Haines (2010). In cases of extreme discrepancy at the county level, the counts from the full count data were compared to published minor civil division data to explore potential instances of missing data, duplicated records, and incorrectly transcribed geographic information. During this process the geographic codes and transcriptions were standardized and classified by state, county, township, place, ward, and enumeration district. IPUMS further enhanced the data by determining urban/rural status, metropolitan areas/districts (Gardner 1998), state economic areas, and census regions.

Table 1 shows the number of counties with too many or too few cases relative to the published census returns. IPUMS processing substantially reduced the number of overcount counties, but naturally had no impact on the undercounting ones. Most of the counties with

significant percentages of over/undercounts are small counties with less than 1,000 people. Counties with errors greater than 5% were generally located in Texas or the U.S. West. There were also a few missing counties in each census year, but these were typically smaller counties; the largest in 1860 and 1870 were Tarrant County, Texas (5,000 people) and Rio Virgin County, Nevada (less than 500 people) respectively. At the national level, after processing, the nineteenth-century full count datasets include slightly fewer cases than were reported in the published volumes (the bottom panel of Table 1). Thus, apart from some scattered localities, the microdata can serve as an accurate source for calculating population denominators for other data sources.

Table 1: Microdata Reporting Rates Compared to Published Counts, 1850-1880

	Error Rate	Data Stage	1850	1860	1870	1880
Over-reporting Counties	> 1%	Original Data Delivery	*	236	119	*
		Post IPUMS Processing	35	62	34	45
	> 5%	Original Data Delivery	*	42	78	*
		Post IPUMS Processing	6	21	15	28
Under-reporting Counties	>1%	Original Data Delivery	20	57	27	77
	>5%	Original Data Delivery	7	15	11	45
Missing Counties	Partial	Original Data Delivery	1	5	4	9
	Total	Original Data Delivery	1	20	13	13
Population Overreported after Processing			10,983	12,246	21,839	65,274
Population Underreported after Processing			16,022	53,213	63,973	77,372
Total Population Underreporting			-5,039	-40,967	-42,134	-12,098
Total Population Underreporting Rate			-0.02%	-0.15%	-0.11%	-0.02%

Source: Haines 2010; S. Ruggles et al. 2024.

*1850 and 1880 do not have Original Data Delivery information for over-reporting because the Ancestry data was merged to our previous full-count data and added no new records

The second significant data processing task was determining household groups. Because of the number of persons and households in the data, manually identifying households was not plausible, requiring an automated process. Each census contained different sets of variables to identify potential households, but the general approach was similar. A family/dwelling number or

surname change generally indicated a new household, but transcription errors and the incidence of complex households containing persons with multiple surnames necessitated a holistic strategy. Other variables such as wealth and occupation, which were often reported only for the household head, were also utilized. The direct identification of household heads by census respondents was a particularly strong indicator of the beginning of a new household, but this information was unique to the 1880 census. Data prior to 1880 include imputed relationship to the head of household (see Table 2 below), but this constructed variable was an output of the household identification procedure, not an input to aid in the delineation process.

Encoding variables within data dictionaries was the third processing step. Data dictionaries are correspondence tables that list each unique string response that occurs in the input data for a particular variable. A second field in the dictionary contains the numeric code IPUMS assigns to that string in the output data. By having one entry that assigns codes to every case that contains the string, data dictionaries are far more efficient than coding each record manually in the data. One of the drawbacks of data dictionaries, however, is the inability to interpret the string differently depending on the specific context for particular cases. As an example, because occupations and wealth are coded in separate dictionaries, outlier cases such as miners with high wealth are not individually identifiable as such (described more below).

Assigning codes to strings within data dictionaries required a different approach in the full count data than in the smaller IPUMS samples, where manually coding the strings was the general practice. For example, there are 1.1 million unique occupation strings and 100,000 unique birthplace strings across all four datasets—too many to code by hand, given budget limitations. Instead, we started from the previous data dictionaries created for the IPUMS samples and applied Jaro-Winkler similarity comparisons (a distance metric used to identify the similarity between two

strings), to assign codes to comparable responses. Finally, any remaining high-frequency strings were manually coded. At the end of these steps, less than 1% of individuals in the full count 1850-1880 censuses had an uncoded occupation, and less than 3% had an uncoded birthplace. These remaining missing values would eventually be allocated, as described below.

Finally, when systemic transcription issues were identified, IPUMS performed manual image lookups and devised fixes. Lookups were sometimes necessary for corrupted data, enumerator idiosyncrasies, and poor data transcription. The most common fixes were for systemic data transcription errors, such as incorrectly interpreting a comma or a decimal in the real and personal estate wealth responses, which led to some inaccurately very wealthy (and very poor) persons. For example, the average wealth of men with the occupation of "sawyer" was significantly higher than men with similar occupations. Manual checks indicated that the occupation of many of the wealthiest "sawyers" in the dataset were in fact incorrectly transcribed "lawyers." In some cases enumerators failed to follow instructions. Enumerator idiosyncrasies included such infelicities as reporting occupations for individuals below the minimum age or interpreting the census questions for cannot read or write as *can* read or write. In some cases, the enumerator crossed out the "not" at the top of the census form to note that their check mark in the field indicated the respondents' literacy, not their illiteracy. Ancestry's contracted data entry operators generally failed to record these divergences from the norm.

The use of the check marks on the census forms for the school attendance, "cannot read", and "cannot write" fields provided particular scope for data capture errors. Stray marks by the original enumerators, crossed out responses by census clerks, and poor image quality from microfilming and digitization all conspired to produce stray marks that data entry operators were prone to record as affirmative responses. We inspected and cleaned errant check mark variables to

the extent feasible, but because of the sheer amount of data we were largely constrained to imperfect programmatic solutions, such as reversing the responses wholesale for selected localities.

To assess data quality, IPUMS compared the published 1870 county-level literacy counts (Haines 2010) to the full count microdata. There are overcounts of illiterate persons over the age of 10 in the Midwest and Northeast in the microdata relative to the published volumes, with many counties overcounting by more than 25%. It is difficult to assess, however, because of uncertainty about how the Census Bureau tabulated illiteracy in 1870 and longstanding concerns about poor enumeration of these census questions (DeBow 1853). Further, matching 1870 sample literacy data to the full count database showed 96.8% agreement between the two sources at the individual level (Ruggles et al. 2023, Ruggles et al 2024). We hope to revisit these data in the future and develop methods to identify further problematic cases.

Another example of a poorly transcribed "check mark" variable is persons denied the vote in 1870. Historical reports counted approximately 40,000 persons denied the vote in the United States in 1870, but the full count microdata shows 125,000 such cases (Giesberg 2021). The IPUMS 1870 1.2% oversample indicates 33,000 persons denied the vote when weighted. As discussed by Oberly (2022, 2024), much of this difference likely resulted from Ancestry's data entry partners ignoring information in the census' margins that indicated a revised entry and from their entering a crossed-out mark as an affirmative (i.e. check mark) response.

When the "check mark" variables (school attendance, literacy, parental nativity, citizenship status, and denied the vote) from the full count data were matched to the 1870 1.2% oversample, they agreed more than 97% of the time. However, the precision of these variables varied. If we (generously) assume the sample data entries are correct, we find that the precision for the check

mark variables ranges between 0.92-0.99, except for "denied the vote," where the precision is only 0.247 for reasons discussed above. But even high correspondence between the sample and full count microdata does not guarantee the data are accurate. Because of concern about the check mark fields expressed by contemporaries (DeBow 1853) and known data transcription issues for the check mark, crime, and disability variables (see Table 2 below for full list of check mark variables), we encourage users to compare full count results to the corresponding IPUMS sample data, published county tabulations, and investigating outliers in the full count data.

Once the data were encoded, the full count files underwent final steps to harmonize them to the broader IPUMS data collection, allocate missing values, construct additional variables, develop integrated documentation, and load the files and metadata into the dissemination system for public distribution. Cases with missing codes were allocated values from similar donor individuals using conventional "hot-decking" procedures documented on the IPUMS website. All cases so altered are indicated by a quality flag. IPUMS imputes relationship-to-head information for the 1850-1870 censuses, which did not include this question on the census forms (Ruggles 1995; Hacker et al. 1999). As with other IPUMS datasets, constructed variables include those identifying the location of an individual's parents and spouse in the household, the number of own children and own children under age 5 in the household, the total number of household members, Hispanic origin, and group quarters status. We also include a variable indicating slave owners who are linked to a slave holding in the slave censuses for 1850 and 1860.

Table 2 shows a partial list of variables available in each of the four full count datasets. In total, the full count population datasets document over 130 million individuals. The 1850 and 1860 datasets include only the nation's free population. The 1870 dataset is the first full documentation

Table 2: Selected Variable Availability for IPUMS Full Count Population Data, 1850-1880

Household Record	1850	1860	1870	1880
Geographic Variables				
State	X	X	X	X
County	X	X	X	X
Urban/rural status	C	C	C	C
Place/City[1]	X	X	X	X
Enumeration district				X
Post Office		R	R	
Street Address				R
Household/Group Quarters Type				
Number of persons in household	C	C	C	C
Multifamily dwelling	X	X	X	X
Household composition	C	C	C	X
GQ quarters residence/Institution type	C	C	C	C
Slaves in Household	X			
Person record				
Family Interrelationships				
Relationship to head	C	C	C	X
Family Pointers (Father, Mother, Spouse)	C	C	C	C
Number of own children, own children under 5	C	C	C	C
Age of eldest and youngest own child	C	C	C	C
Demographic Characteristics				
Age in years, months persons under one	X	X	X	X
Sex	X	X	X	X
Race	X	X	X	X
Hispanic	C	C	C	C
Month of marriage			X	
Married within year	✓	✓	✓	✓
Marital Status	C	C	C	X
Birthplace (state or country)	X	X	X	X
Parental Birthplace				X
Parent foreign born			✓	
Citizenship			✓	
Voting rights denied			✓	
Economic Status, Education				
Labor force status	C	C	C	C
Occupation[2]	X	X	X	X
Industry -- 1950 basis	X	X	X	X
Months Unemployed				X
Value of real property	X	X	X	
Value of personal property		X	X	
School attendance	✓	✓	✓	✓
Literacy	✓	✓	✓	✓
Other				
Deaf and Dumb, Blind, Insane, Idiotic	inc	inc	inc	✓
Pauper, Crime committed	inc	inc		
Sickness				X
Surname, Given Name	R	R	R	R
Number of slaves linked to owner	X	X		
N	19,981,435	27,451,053	38,515,225	50,140,482

Notes: Variables denoted by "X" are census questions available in a given year, while variables denoted by "C" were constructed by IPUMS using logical rules. "R" indicates variables that are only available in restricted versions of the full count data. "✓" indicate variables where we suspect data transcription issues in the checkmark variables to be used with caution, and variables denoted by "inc" are variables we suspect incomplete data transcriptions and should be used with caution. Other variables not shown include population of urban places, state economic area, standard metropolitan area, metropolitan district, institutional support, and Spanish surname.

[1]Cities with populations greater than 10,000 are identified. Smaller places can be identified in the restricted data

[2]Occupational Classification is based on the 1880 and 1950 Census Bureau Classification codes. HISCO classification codes are also available for 1880

of the nation's Black population, five years after the general emancipation in 1865, and serves as the earliest baseline for changes in the Black population. Some variables such as names are only available in restricted versions of the dataset. Users requiring these variables should contact IPUMS at ipums@umn.edu for more information.

Discussion and Conclusion

No census—or census dataset—is perfect. Estimates suggest that census coverage errors were high in the nineteenth century and that the administration of the census faced many challenges. By 1880, however, significant reforms were in place and undercounts were significantly lower (Hacker 2013; King and Magnuson 1995). In this paper we discuss the administration of the census in the nineteenth century and new IPUMS full count datasets of the 1850, 1860, 1870, and 1880 censuses based on the copies of the original manuscript returns. The enumeration and administrative processes of the nineteenth century impact the IPUMS full count datasets in ways that informed researchers should know and understand.

These datasets provide researchers with improved resources to study socioeconomic and demographic outcomes in the United States and open many new research possibilities. Because the data are complete, researchers can study small subpopulations and sub-county geographies. Researchers can also calculate custom aggregated summary statistics, including measures of residential segregation (e.g. Logan and Parman 2017; Grigoryeva and Ruef 2015) and economic opportunity (Long and Ferrie 2018; Song et al. 2019). Because the nineteenth century Census Office tabulated results by hand, little beyond basic population counts was published for sub-county areas, and only basic counts of broad categories were tabulated for counties. The few statistics such as illiteracy that were provided are often incompatible with those of later census years. The IPUMS full count datasets allow researchers to create aggregated summary files for

sub-county areas. Numerous other applications are possible, including analyses of slaveholding, slave population processes, segregation, and correlates of morbidity and mortality.

There are several concerns, however, stemming from the original enumeration and the processing challenges posed by these large databases. In contrast to data entry for the low-density IPUMS samples (Ruggles et al. 2020), which was conducted at the University of Minnesota, data entry for the full count datasets was contracted out by Ancestry.com, primarily by Beijing Formax, and conducted without MPC training, oversight, verification, and guidance. Given the sheer amount of data, some string values for variables such as occupation and birthplace were coded with automated methods, and manual verification of all the data was not possible. Because of the quality of these original enumerations and known data transcription issues, data quality is likely higher in the sample datasets than in the full count data transcriptions for the "check mark" variables such as denied the vote. Users may consider using IPUMS sample data in conjunction with the full count data for national level analyses and compare the results. Using the 1850-1880 full count datasets is recommended for analyses requiring larger case counts such as studies of local areas, applying local area fixed effects, studies of smaller demographic groups, the construction of contextual variables, and linking other historical data to the IPUMS full count census data. With appropriate caution, these data open a host of opportunities to analyze U.S. population, economic, and social history.

Acknowledgements

Data Availability: The IPUMS Ancestry Full Count Population Census Data is available for download from usa.ipums.org. Users interested in restricted versions of the data should contact ipums@umn.edu for more information. The DOI for the IPUMS Ancestry Full Count Census Data is https://doi.org/10.18128/D014.V4.0

Corresponding Author: Matt A. Nelson, nels5091@umn.edu

Disclosure Statement: The authors report there are no competing interests to declare.

References

Abramitzky, R., L. Boustan, and K. Eriksson. 2021. "Automated linking of historical data," *Journal of Economic Literature* 59(3): 865-918.

Anderson, M.J. 2015 2d. *The American Census: A Social History*. Yale University Press.

Blake, K. 1996. "'First in the Path of the Firemen': The Fate of the 1890 Population Census," *Prologue Magazine* 28(1): 64–81.

DeBow, J.D.B. 1853. *The Seventh Census of the United States: 1850*. Washington: Robert Armstrong.

DeBow, J.D.B. 1854. *Compendium of the Seventh Census*. Washington: Beverley Tucker, Senate Printer.

Curti, M. 1959. *The Making of an American Community: A Case Study of Democracy in a Frontier County*. Stanford, California: Stanford University Press.

Dorman, R.L. 2008. "The Creation and Destruction of the 1890 Federal Census," *The American Archivist* 71(2): 350-83.

Dreiser, L.H. 1944. "Microfilm Saves Expensive Storage," *Domestic Commerce* 32(8): 13 and 24.

Fogel, R. W. and S. L. Engerman. 1974. *Time on the Cross: The Economics of American Negro Slavery*. Boston: Boston, Little, Brown.

Gardner, T.K. 1998. *The Metropolitan Fringe: Suburbanization in the United States before World War II*. Ph.D. Dissertation, University of Minnesota.

Genadek, K.R. and J. T. Alexander. 2022. "The Missing Link: Data Capture Technology and the Making of a Longitudinal U.S. Census Infrastructure," *IEEE Annals of the History of Computing* 44(4): 57-66. DOI: 10.1109/MAHC.2022.3195001

Giesberg, J. 2021. "'A Muster-Roll of the American People': The 1870 Census, Voting Rights, and the Postwar South," *Journal of Southern History* 87(1): 35-66. https://doi-org.ezp1.lib.umn.edu/10.1353/soh.2021.0001

Goeken, R., L. Huynh, T.A. Lynch, and R. Vick. 2011. "New Methods of Census Record Linking," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44(1): 7-14. https://doi-org.ezp1.lib.umn.edu/10.1080/01615440.2010.517152

Goeken, R., C. Nguyen, S. Ruggles, and W. Sargent. 2003. "The 1880 U.S. Population Database," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 36(1): 27-34. https://doi-org.ezp1.lib.umn.edu/10.1080/01615440309601212

Grigoryeva, A. and M. Ruef. 2015. "The historical demography of racial segregation," *American Sociological Review* 80(4): 814-842.

Hacker, J.D. 2003. "Rethinking the 'early' decline of marital fertility in the United States," *Demography* 40(4): 605-620.

Hacker, J.D. 2008. "Economic, Demographic, and Anthropometric Correlates of First Marriage in the Mid Nineteenth Century United States," *Social Science History* 32(3): 307-345.

Hacker, J.D. 2011. "A census-based count of the Civil War dead," *Civil War History* 57(4): 307-348.

Hacker, J.D. 2013. "New Estimates of Census Coverage in the United States, 1850-1930," *Social Science History* 37(1): 71-101.

Hacker, J. D., S. Ruggles, A.R. Foroughi, B.D. Jarvis, and W.L. Sargent. 1999. "Public use microdata samples of the 1860 and 1870 US Censuses of population," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 32(3): 125-133. https://doi-org.ezp1.lib.umn.edu/10.1080/01615449909598933

Hacker, J. D. and E. Roberts. 2017. "The impact of kin availability, parental religiosity, and nativity on fertility differentials in the late 19th-century United States," *Demographic Research* 37(34): 1049.

Haines, M.R., and Inter-university Consortium for Political and Social Research. 2010. Historical, Demographic, Economic, and Social Data: The United States, 1790-2002. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR02896.v3

Hall, P. K. and S. Ruggles. 2004. "Restless in the Midst of their Prosperity': New Evidence of the Internal Migration Patterns of Americans, 1850-1990," *Journal of American History* 91(3): 829-846.

Helgertz, J., J. Price, J. Wellington, K.J. Thompson, S. Ruggles, and C.A. Fitch. 2022. "A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55(1): 12-29. https://doi-org.ezp1.lib.umn.edu/10.1080/01615440.2021.1985027

Karbeah, J. M. and J.D. Hacker. 2023. "Racial residential segregation and child mortality in the southern United States at the turn of the 20th century," *Population, Space and Place* 29(6): e2678.

Kennedy, J.C.G. 1860. "Report of the Superintendent of the Census, November 30, 1860" in *Report of the Secretary of the Interior*, S. Exdoc. 36-1.

Kennedy, J.C.G. 1864: *Population of the United States in 1860; Compiled from the Original Returns of the Eighth Census*. Washington: GPO.

King, M.L. and D. L. Magnuson. 1993. "Appendix B, A Procedural History of the Enumeration Process in the 1880 U.S. Census," *User's Guide and Technical Documentation 1880 PUMS*. University of Minnesota: Social Research Laboratory.

King, M.L. and D.L. Magnuson. 1995. "Perspective on Historical U.S. Census Undercounts," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 19(4): 455-466. https://doi.org/10.2307/1171475

Lathrop, Barnes F. 1948. "History from the Census Returns," *The Southwestern Historical Quarterly* 51(4): 293-312.

Logan, T. D. and J. M. Parman. 2017. "The national rise in residential segregation," *The Journal of Economic History* 77(1): 127-170. https://doi.org/10.1017/S0022050717000079

Long, J. and Ferrie, J. (2018). "Grandfathers matter (ed): occupational mobility across three generations in the US and Britain, 1850–1911," *The Economic Journal* 128(612): F422-F445.

Magnuson, D.L. 1995. *The Making of a Modern Census: The United States Census of Population, 1790-1940*. Ph.D. Dissertation, University of Minnesota.

Magnuson, D.L. and M. L. King. 1995. "Comparability of the Public Use Microdata Samples: Enumeration Procedures," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 28(1): 27-32. https://doi-org.ezp1.lib.umn.edu/10.1080/01615440.1995.9955310

Munroe, J. P. 1923. "The Census of 1880," *A Life of Francis Amasa Walker*. New York: H. Holt & Co.

Nelson, M. A. 2020. "The decline of patrilineal kin propinquity in the United States, 1790–1940," *Demographic Research* 43(18), 501-532.

Oberly, J.W. 2022. "'When Homer Nods' in Tabulating Question 20 on the 1870 Census, Denial of the Right to Vote: Errors Made in the Manuscript Census, Repeated by Ancestry.com, and Coded into the IPUMS 1870 Full Count File." [Manuscript submitted for publication].

Oberly, J.W. 2024. "Voter Eligibility, Voter Turnout, Partisan Votes, and the Election of 1870 in Maine's Counties and Towns," *Maine History* 56(1): 28-49.

Rosenbloom, J. L. and W. A. Sundstrom. 2004. "The decline and rise of interstate migration in the United States: Evidence from the IPUMS, 1850–1990," *Research in Economic History*, Emerald Group Publishing Limited 22: 289-325.

Ruggles, S. 1994. "The transformation of American family structure," *American Historical Review* 99(1): 103-128. https://doi-org.ezp3.lib.umn.edu/10.2307/2166164

Ruggles, S. 1995. "Family interrelationships," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 28(1): 52-58. https://doi-org.ezp1.lib.umn.edu/10.1080/01615440.1995.9955314

Ruggles, S. 1997. "The Rise of Divorce and Separation in the United States, 1880-1990." *Demography* 34(4): 455-466. https://doi-org.ezp3.lib.umn.edu/10.2307/3038300

Ruggles, S. 2023. 'Collaborations Between IPUMS and Genealogical Organizations, 1999-2022," *Historical Life Course Studies* 13:1-10. DOI: https://doi.org/10.51964/hlcs12920

Ruggles, S., S. Flood, M. Sobek, D. Backman, A. Chen, G. Cooper, S. Richards, R. Rogers, and M. Schouweiler. 2023. IPUMS USA: Version 14.0 [dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V14.0

Ruggles, S. and D. L. Magnuson. 2020. "Census Technology, Politics, and Institutional Change, 1790-2020," *Journal of American History* 107(1):19-51. https://doi.org/10.1093/jahist/jaaa007

Ruggles, S. and R. R. Menard. 1990. "A public use sample of the 1880 US census of population," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 23(3): 104-115. https://doi-org.ezp1.lib.umn.edu/10.1080/01615440.1990.10594201

Ruggles, S., M.A. Nelson, M. Sobek, C.A. Fitch, R. Goeken, J.D. Hacker, E. Roberts, and J.R. Warren. 2024. IPUMS Ancestry Full Count Data: Version 4.0 [dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D014.V4.0

Sobek, M. 2001. "New Statistics on the U.S. Labor Force, 1850-1990," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 34(2): 71-87. https://doi-org.ezp1.lib.umn.edu/10.1080/01615440109598973

Song, X., C.G. Massey, K.A. Rolf, J.P. Ferrie, J.L. Rothbaum, and Y. Xie. 2020. "Long-term decline in intergenerational mobility in the United States since the 1850s." *Proceedings of the National Academy of Sciences*, *117*(1), 251-258.

Thernstrom, S. 1964. *Poverty and progress: Social mobility in a nineteenth century city*. Harvard University Press.

Thernstrom, S. and P. R. Knights. 1970. "Men in motion: Some data and speculations about urban population mobility in nineteenth-century America," *The Journal of Interdisciplinary History* 1(1): 7-35.

U.S. Census Office. 1860. "Instructions to U.S. Marshals, Instructions to Assistants." Washington: GPO.

U.S. Census Office. 1870. "Instructions to Assistant Marshals." Washington: GPO.

U.S. Census Bureau. 2002. *Measuring America: The Decennial Censuses from 1790 to 2000*. Prepared by J. G. Gauthier. Washington, D.C.: U.S. Department of Commerce.

Walker, F.A. 1878. *Report of the Superintendent of the Census*. Washington: GPO.

Walker, F.A. 1883. *Compendium of the Tenth Census, June 1, 1880.* Washington: GPO.

Wildsmith, E., M.P. Gutmann, and B. Gratton. 2003. "Assimilation and intermarriage for U.S. immigrant groups, 1880-1990," *The History of the Family: An International Quarterly* 8(4): 563-584. https://doi-org.ezp3.lib.umn.edu/10.1016/j.hisfam.2003.05.001

Wright, C.D. and W.C. Hunt. 1900. *The History and Growth of the United States Census*. Washington: GPO.

9 Stat. 395 1849 (Statutes at Large, volume 9, page 395; Secretary of the Interior given supervision of the United States census)

9 Stat. 402 1849 (Statutes at Large, volume 9, page 402; Census Board created to prepare for taking the seventh census)

9 Stat. 428 1850 (Statutes at Large, volume 9, page 428; Provisions for taking the seventh and subsequent censuses)

16 Stat. 118 1870 (Statutes at Large, volume 16, page 118; Amending Act of May 23, 1850 in regard to time of making returns of enumeration)

20 Stat. 473 1879 (Statutes at Large, volume 20, page 473; Provisions for taking the tenth and subsequent censuses)

H.R. No. 3 1870 (Report of the Committee on the Ninth Census)