IPUMS

Working Papers

Historical Context and Creation of the IPUMS Ancestry Full Count U.S. Population Census Data 1900-1930

Matt A. Nelson†
University of Minnesota

Diana L. Magnuson
University of Minnesota

Steven Ruggles
University of Minnesota

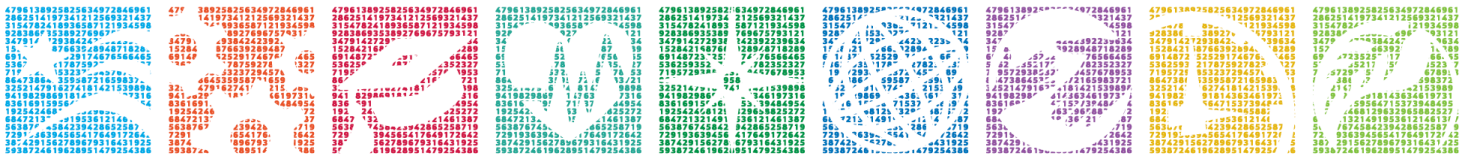
Matthew Sobek
University of Minnesota

Lap Huynh
University of Minnesota

June 2024

Working Paper No. 2024-04
DOI: https://doi.org/10.18128/IPUMS2024-04

†Address correspondence to Matt Nelson: nels5091@umn.edu. This project was supported in part by funding from the Minnesota Population Center (P2CHD041023) and by a grant from the Eunice Kennedy Shriver National Institute for Child Health and Human Development (R01-HD078322).



Historical Context and Creation of the IPUMS Ancestry Full Count Population Census Data 1900-1930

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)
Steven Ruggles, *University of Minnesota* (0000-0001-5353-2578, ruggles@umn.edu)
Matthew Sobek, *University of Minnesota (sobek@umn.edu)*
Lap Huynh, *University of Minnesota (huyn0064@umn.edu)*

Abstract:

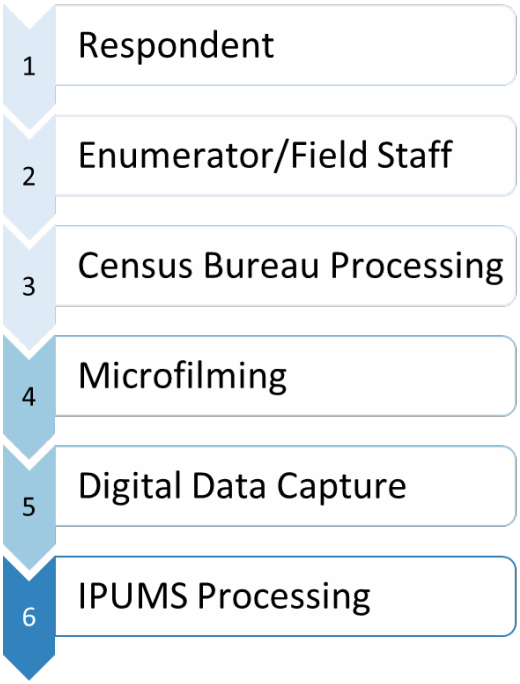
IPUMS recently released final versions of full count census data for the United States 1900-1930. The information contained in these files is the product of three broad work stages: historical census enumeration, digitization, and IPUMS processing. The data were produced within an evolving institutional context and subjected to subsequent processes that had important ramifications on the final product. This paper documents these histories and processes and their implications for research. Because of the datasets' sheer size and scale, the development of these files necessitated applying different methods and approaches to assess data quality and correct the data. We document cases where data quality was affected not only by choices made by the Census historically, but also by data transcription errors in the modern day. Finally, we describe our approaches to processing the data, and we note some of the implications for research these various decisions have. As with any dataset, researchers should use this resource critically for their particular research questions and consider the data creation process from respondent to digital dataset. Despite some limitations and liabilities, the IPUMS full count data provide a powerful and valuable resource to study demographic effects on a variety of health and socioeconomic questions.

Keywords: census, microdata, population, demography

Introduction

The IPUMS Full Count Census data collection is the largest publicly accessible population research database available for any country. This paper documents the historical census enumeration, subsequent digitization of the census forms, and modern data processing that produced this unique resource for 1900-1930. Figure 1 depicts the stages through which the data have passed to reach present-day researchers. Historical data capture entails what information a respondent provided for a household, the work of enumerators and field staff to record the information on the census forms, and the processing by the Census Bureau to produce published statistics. Digitization was achieved via microfilming and subsequent data transcription by genealogical companies. Finally, these transcribed data were shared with IPUMS, which processed and harmonized the information to enable analysis with other U.S. censuses and surveys.

Figure 1: Flowchart of Information from Respondent to IPUMS Data



As Figure 1 demonstrates, researchers are at least six steps removed from the original information collected a century or more ago. Each step of the process raises the prospect of errors, some of which are irreversible. For example, if the respondent for a household provided incorrect information, modern-day analysts have limited means for making corrections, assuming they even detect the error. This paper aims to provide an overview of these issues and their known implications for future research. Despite some imperfections in the data, the content and scope of this data collection allows analyses of previously understudied groups. Because of their size, the databases contain fine geographic detail for contextual and local studies. These demographic and socioeconomic data offer the potential for linking to other surveys and historical collections, enabling new and innovative research.

Historical Census Enumeration

Censuses in the early twentieth century were conducted door-to-door by large numbers of temporary enumerators. The data are the product of millions of interpersonal interactions that recorded dozens of information items onto paper forms. Census procedures themselves were not static. The government office charged with conducting the census evolved considerably over this period and continually refined its methods, the questionnaire, and the instructions to field workers, in an effort to improve results.

The census of 1900 was taken under the Census Act of 1899 (Department of Interior 1900a). Census day was June 1, 1900, and enumerators had until July 1, 1900, to complete their returns and forward them to their respective supervisors. Each enumerator was responsible for canvassing an enumeration district that was not to exceed 4,000 inhabitants. In cities with 8,000 or more inhabitants, the enumeration was to be completed within two weeks of June 1, 1900.The process of acquiring the temporary 60,000-person workforce was refined in 1900 by requiring

potential enumerators to take a written exam to screen their suitability for the work. Granted, this was a take-home test, but the provision did signal to applicants that there were basic requirements for the job. The test schedule was an exact copy of the population schedule, mailed to each candidate, to be "filled out in hypothetical manner" using a sample narrative (Annual Report 1900, 293; Department of Interior 1900b). Candidates returned the completed schedule to their supervisor with certification that they had not received any assistance filling out the form. Enumerators in 1900 received a 64-page instruction booklet delineating everything from general instructions (the Census Act, care of schedules, enumerator's rights, use of the telegraph) to special instructions regarding the schedules of population, agriculture, and manufactures (Department of Interior 1900b).

An important procedural innovation of the 1900 Twelfth Decennial Census was the introduction of the "street book." The 3.5-inch x 8-inch street book was designed to support thorough canvassing of city populations, by accounting for every building or vacant space within a given enumeration district. The 1910 street book contained 63 preprinted pages, including seven fields (name of street, house number, flat or room number, description, date of visit, remarks, and date collected) for enumerators to note where they were unsuccessful in securing information at the first visit ("Twelfth Census of the United States, Enumerator's Street Book;" Thirteenth Census of the United States, Enumerator's Street Book"). This tool provided a new method for the district supervisors and, by extension, the Census Office, to monitor the "completeness and correctness" of the returns and, "so far as possible," to avoid "complaint as to the correctness of the population returns." The street book aided the enumerator in accounting "for each and every house, building, or place of abode, of whatever kind, within the limits of his enumeration district, the record being made in such form as to permit of easy verification of the completeness of the house-to-house

canvass" (U.S. Census Office 1900, 6-8; Magnuson 1995, 177-178; U.S. Census Bureau 2002, 34).

The Census Office became the permanent Census Bureau in 1902 with the passage of "An Act to provide for a permanent Census Office" (U.S. Census Office 1902). The case for continuous administration was first made in 1854 when Superintendent James D.B. DeBow included a section titled "The Office" in his *Compendium of the Seventh Census*:

> Unless there is machinery in advance at the seat of Government no census can ever be properly taken and published … Each census has taken care of itself … In Washington, as soon as an office acquires familiarity with statistics, and is educated to accuracy and activity, it is disbanded, and even the best qualified employee is suffered to depart (DeBow 1853 18).

Superintendent Robert Porter described the "obstacles confronting him" in 1890 when he took up the work:

> When I was appointed I had nothing but one clerk and a messenger, and a desk with some white paper on it … Then the difficulty comes in getting your force together, picking out your men. I was not able to get more than three of the old men from this city ... Then, knowing all the old special agents of the Tenth Census, I wrote asking them if they were prepared to take up the work again. Some were and some declined … Some of them were dead and some in private business. I succeeded in getting one from Colorado … With these men we started up the organization (Holt 1929, 27).

The enormity of the task confronting the Census Office grew steadily with each succeeding decennial census. To successfully conduct the administrative work, oversee the hiring and training of field staff, coordinate data capture and tabulation, and complete a myriad of other tasks, the Census Office needed to be a permanent agency. The case for a permanent census office that reasonably distributed the work throughout each decade, built on institutional capital, made data readily available, and did it all more cost effectively is historically well documented (Wright and Hunt 1900, 79-83; "Permanent Census Office" 1901; Cummings 1913; Wilcox 1914; Eckler

1972). The establishment of a permanent Census Bureau thus counts as a key turning point in the coverage and process of the U.S. census work.

With the establishment of the permanent Census Office, the "shape" of the emerging federal statistical system was in place, but the precise mechanisms for implementing that system was worked out over the period 1910-1930 (Anderson 2010). In the context of building the infrastructure for continuous statistical data collection, the Census Bureau was both developing administrative structures for the decennial censuses and constructing "population policies needed for an urban industrial society" (Anderson 2015). Socially, politically, and economically, the United States in these decades was still a product of the nineteenth century, and this legacy extended to the early Census Bureau.

The 1910 census, enumerated under the Thirteenth Decennial Census Act, was the first to be taken under the oversight of a permanent Census Bureau ("An Act to provide for the Thirteenth and subsequent decennial censuses"). The census administration in Washington looked forward to having ample time to prepare for and expand the recruitment, organization, and training of supervisors, enumerators, and other field staff (U.S. Bureau of the Census 1907, 18-19). In a compromise with civil service reformers, in-person examinations for would-be enumerators were proctored at locations across the country, replacing the take-home tests piloted in 1900 (Magnuson 1995, 139-140). Supervisors were "expressly instructed" not to allow partisan politics to play a role in the selection of enumerators, but it was generally acknowledged that the practice was not uncommon. Despite their limitations for determining the "character and efficiency" of individuals applying for enumerator positions, mass examinations offered at least a minimal screening mechanism for hiring the enormous number of temporary employees needed for the decennial census field work (U.S. Bureau of the Census 1908b, 20-22; U.S. Bureau of the Census 1911, 7-

9; Durand 1910, 54; Magnuson 1995, 140-142). In addition to "very full written and printed instructions" issued to each supervisor, conventions were held to train supervisors "in convenient cities in different parts of the country" (U.S. Bureau of the Census 1911, 6; U.S. Bureau of the Census 1910). At these conventions, the Director or Assistant Director of the Census, accompanied by the chief statistician for population or the chief statistician for agriculture, gave oral instructions to supervisors and answered questions. Determining supervisors' districts and filling the supervisor posts proceeded at a brisk pace.

The decennial difficulty in training and supervising a nationally dispersed and temporarily employed enumerator force, numbering over 71,000 in 1910, was clear. In 1910, supervisors in large cities had support staff in the field to provide "continuous personal supervision and instruction" to enumerators. These "inspectors" examined enumerators' work on an almost daily basis, and the inspectors' tasks included "answering such questions as arose from time to time, checking the work of the enumerators at random, and otherwise assisting and directing them" (U.S. Bureau of the Census 1911, 12). In rural areas where no inspectors were installed, enumerators were expected to send a copy of part of their first day's work to their supervisor through the U.S. mail. Supervisors examined the schedules, corrected them if necessary, and returned them to enumerators (Magnuson 1995, 182; U.S. Census Bureau 2002, 45-46). This back-and-forth communication and oversight between supervisors and enumerators identified fieldwork problems early, rather than only at the end of the enumeration, when it was generally too late to correct procedural errors.

The census reference day in 1910 was moved to April 15th to accommodate vacationing urban dwellers (1910 Census Overview). The Director of the Census argued that "The habits of the American people have so changed that it is no longer possible to enumerate the residents of

our large cities on a date as late as June 1 with any accuracy" (U.S. Bureau of the Census 1908a, 23). More than half of the population was urban or "semi-urban," and "the difficulties attending the count of the summer absentees from urban homes can no longer be overcome by ordinary expedients, such as the "prior-schedule" and "resort to the mails" (U.S. Bureau of the Census 1908a, 23). The Census Act thus stated that it was the duty of each enumerator to commence enumeration on April 15th, "unless the Director of the Census in his discretion shall defer the enumeration in said district by reason of climatic or other conditions which would materially interfere with the proper conduct of the work" (Department of Interior 1909, section 20).

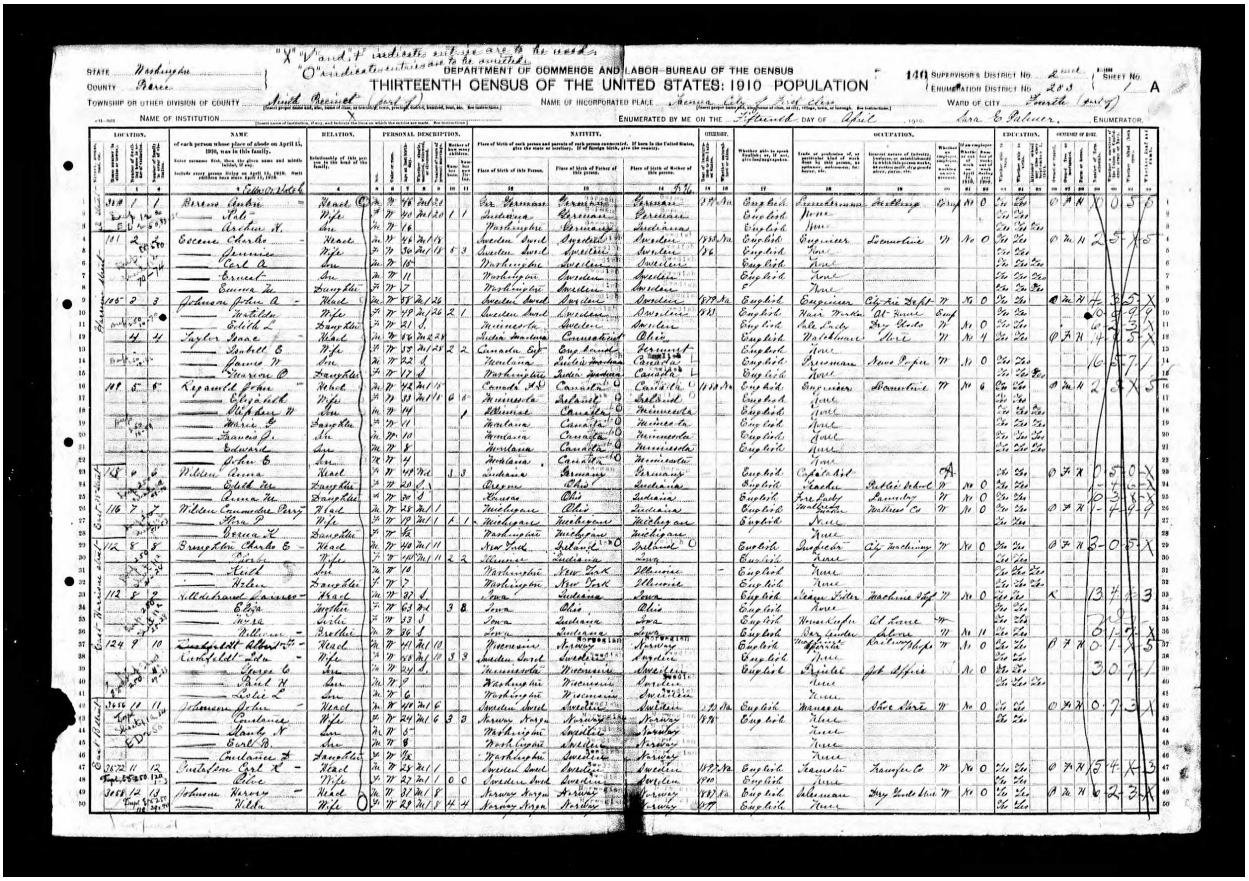
Despite the refinements to field staff training and oversight and the obvious advantage of a permanent Census Bureau, the Director of the Census continued to lament the quality of his enumerators, just as his predecessors had done for decades. Director E. Dana Durand devoted an entire section of his 1912 annual report to the unsatisfactory character of present methods. Durand wrote,

> The term of employment, particularly of enumerators, is so short and the pay so small that it is very difficult to induce competent persons to take the job; and, finally, that there is no adequate means of holding supervisor or enumerators responsible for conscientious and thorough work … An enumerator, once selected, knows that at most the only penalty for unsatisfactory work will be failure to receive his comparatively small compensation, and that in fact it is scarcely likely that his incompetence will be discovered until after he has received his pay (U.S. Bureau of the Census 1913, 21-23).

Durand's dour outlook may have been influenced by fraud allegations related to the 1910 census, which he summarized as follows: "In more than twenty cities the Census Bureau discovered extensive 'padding' of the returns for 1910, in some cases amounting to as much as 30 or 40 percent," and that this "cannot but raise suspicion as to the accuracy of the figures for some other places where the vigilance of the bureau officials may not have succeeded in detecting the fraud"

(Durand 1913). Figure 2 shows an example of fraudulent records. The note at the top of the page indicates that records with an "O" in the relation column are to be omitted, which in this case applies to every line on the page. Unfortunately, such historical subtleties would be lost during digitization decades later, and these records entered the microdata. Such erroneous records had to be identified later where possible and removed during data processing, as noted below.

Figure 2: 1910 Enumerator Sheet of Fraudulent Records



Source: Ancestry.com *2006*; Original data: Thirteenth Census of the United States, 1910.

Durand also had to deal with allegations of undercounting. He qualified his discussion of undercounts in the 1910 census by stating that "It would be impossible for the Census Bureau by any investigation at Washington of the schedules of the enumerators to ascertain cases of undercounting of the population through the carelessness or inefficiency of the enumerators." Supervisors were directed to review schedules as they came in and, before sending the schedules

to Washington, "adopt all available means for perfecting returns." Despite these efforts, it was "impossible" for even "the most efficient supervisor to guarantee the correctness of the enumeration through his entire district" (Durand 1913, 563, 566-567; Merriam 1901; U.S. Bureau of the Census 1911, 23). Notably, to whatever extent such undercounting did occur, there is no modern corrective; that information is simply lost.

The 1920 decennial census of population continued administrative efforts to refine and improve the training and oversight of census field staff. Census day was moved again, this time to January 1, 1920, to accommodate the memories of farmers (harvest figures would be "fresh in their minds"), and more people would be home than in April or June (1920 Census Overview). Schedules, forms, and instructions for enumerators, supervisors, special agents, and inspectors were revised and amplified (U.S. Bureau of the Census 1920a; U.S. Bureau of the Census 1919a; U.S. Census Bureau 2002, 58). Continuing the trend toward greater contact between the Director of the Census and the district supervisors, with an eye toward increasing the accuracy of the census, the volume of correspondence and training of those supervisors increased in 1920 (U.S. Bureau of the Census 1919b; U.S. Bureau of the Census 1919c; U.S. Bureau of the Census 1919d; U.S. Census Bureau 1919a; "Qualifications, Duties, and Compensation of Census Enumerators").[1] The 1910 innovation of supervisors' conferences, directed by the census administration, was continued for the 1920 census of population. Eleven such conferences were held in cities across the United States, which all but 34 of the 372 total supervisors attended (U.S. Bureau of the Census 1920b; "Untitled Press Release 29 November 1919"). Just as the volume of training material increased for supervisors, so too for enumerators. Supervisors were directed to orally instruct enumerators

[1] Supervisors received numerous form letters (fifty in all) and other miscellaneous correspondence. See NARA RG 29,212/181: 25.

whenever possible "in convenient numbers and at convenient points" (U.S. Bureau of the Census 1919c, 3) Inspectors or special agents were again a factor in the oversight of enumerators in 1920. The field staff inspector provided another layer of reporting and training for enumerators. In large cities, a supervisor employed one or two inspectors (also called special agents) for the purposes of supervising, assisting, and instructing enumerators "to make sure that they are performing their duties intelligently, industriously, and faithfully" (U.S. Bureau of the Census, 1919d, 3) Women were also employed as supervisors for the first time; three received original appointments, and two more were later appointed to fill vacancies (Magnuson 2012, 144).

The 1920 Director's report did not mention investigations related to census fraud other than noting the Bureau was monitoring the situation through "a force of from seven to ten clerks" examining suspicious schedules for padding or irregularities in the enumeration. Particular attention was being paid to "booster, tourist, and winter-resort cities." After five months of work, the special examination was completed in 73 cities, but "no serious discrepancies" had been discovered (U.S. Bureau of the Census 1920b, 20).

The training and oversight of field staff at the 1930 census of population continued the trend of refining and scaling up procedures implemented at previous censuses (U.S. Census Bureau 2002, 59-61). Enumerators received a revised pamphlet of instructions, a record book (like the old street book, first instituted in 1900), and oral instructions from their respective supervisors. Beginning in 1930, inspectors were now called "field assistants," but they served the same purpose of interfacing between supervisors and enumerators and instructing and advising enumerators in the field. The 1930 census reference day was moved to April 1, where it has remained since.

Census Bureau Processing

The establishment of a permanent Census Bureau in 1902 ensured that staff expertise was preserved and built upon over the decades. Census data capture and processing—the methods and technologies used by the Census Bureau to transform the manuscript census forms into statistical tables—underwent significant transformation from 1900 to 1930. The technological innovations by Bureau staff were practical responses to processing bottlenecks within the system of tabulation (Ruggles and Magnuson 2020; Truesdell 1965). Machine tabulation via punch cards had begun in 1890, which introduced a coding stage to census processing on the way to producing tables for publication. However, while the coding process for tabulation certainly involved decision-making and the possibility of errors, that work has little bearing on the modern microdata, which are derived only from what was written on the forms themselves.

Instructions for clerks examining population schedules in 1910 emphasized that the most important duty of schedule review was ascertaining that the enumerator covered their entire district, enumerated everyone within the district, and that "the information required under the different column headings has been, on the whole, carefully and accurately secured by the enumerator and intelligently entered." ("Instructions For Clerks Examining Population Schedules to Take Effect June 11;" "Memo and Notes of J.A. Hill;" "Instructions for Editing: Family Card, form 8-3315")

"Editing" was the first step in preparing the schedules for punching clerks who would then create punch cards for machine tabulation (U.S. Bureau of the Census 1911, 42). Editing work for the 1920 census was similarly documented in the 1920 director's report (U.S. Bureau of the Census, 1920b, 20-21). The editing stage ranged from correcting entries for specific fields to striking out entire records deemed redundant. Occupation codes were written in the margin of the

forms by clerks in 1920, and the 1930 census was the first to dedicate columns in the body of the form "for office use only" for clerical coders. Edits to the forms in the census office may or may not have been captured during data transcription in the 21st century, depending on the nature of the edits and where they occurred on the census page.

The Census Bureau produced published reports and statistics for a variety of demographic and socioeconomic topics at different geographic levels. These published volumes were the main product of the Census; the notion of microdata analysis of these records was, of course, decades in the future. The published counts are an important resource for modern processing of the census databases, as they offer some scope for validating (or "ground truthing," in machine learning terms) the microdata classifications and locating errors. It is important to note, however, that we do not know all the steps the Census Bureau took to exclude certain records or how it tabulated results in particular tables (including the punch card coding process itself). Because of these intervening factors, as well as the inherent imperfections of the digitization process, researchers cannot usually expect an exact match of the full count data to published results.

Although it can be tempting to view the published census results as a kind of "gold standard" to which the microdata should aspire, those historical tabulations were sometimes manipulated in ways we may not wish to duplicate. Out of concern for public perceptions about the quality of its work, throughout the early 20th century, the Bureau imposed preconceptions on some of its results via verification procedures designed to weed out suspicious-looking outcomes. For example, women in unusual occupations were examined, flagged, and recoded to more typical jobs, and the scrutiny this entailed likely influenced the clerks' work up and down the line. Minorities in high status occupations drew similar attention (Conk 1981; "List of Cards to be Rejected in the Sort of the Fourteenth Census Occupation Cards for Females and Males"). We can

only speculate about other "improvements" the Bureau made. But most of the census actions during tabulation did not involve editing the census forms themselves, and thus such actions have no direct bearing on the modern microdata.

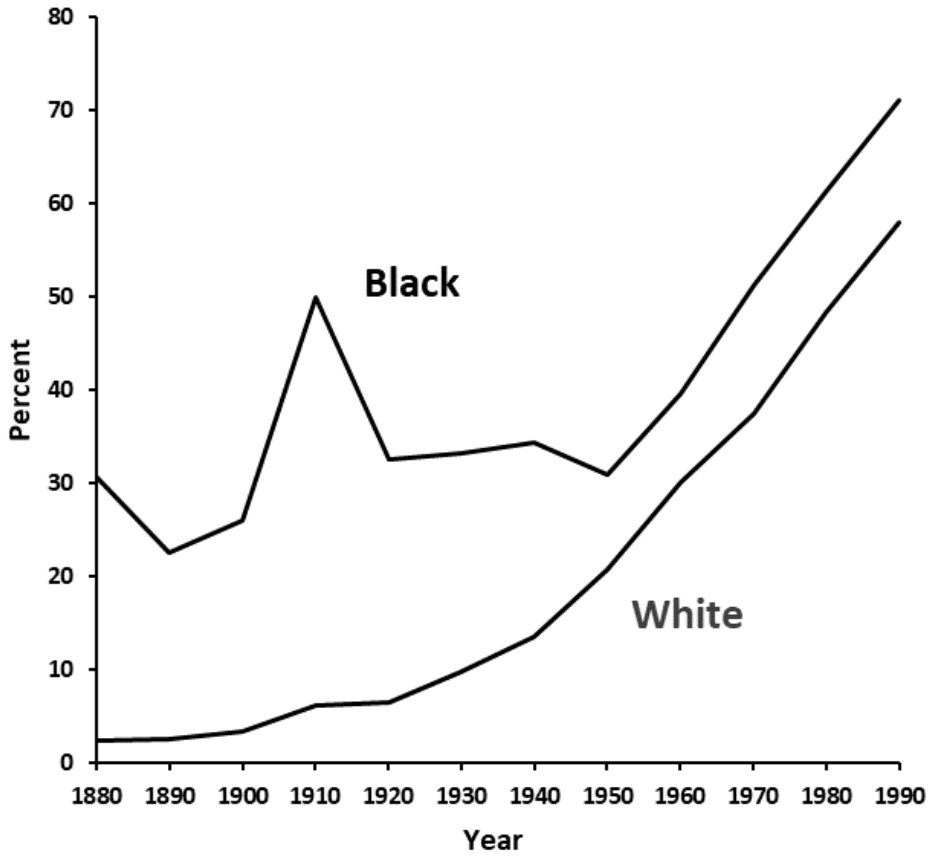
The instructions to enumerators were steadily refined to improve the quality of the returns, but sometimes this had unintended consequences leading to a lack of comparability across census years. In hope of better capturing women's labor in family enterprises, the 1910 census introduced strong language. In the opening paragraph of the 1910 occupation instructions, enumerators were told that the occupation of a woman "is just as important, for census purposes, as the occupation followed by a man. Therefore, it must never be taken for granted, without inquiry, that a woman … has no occupation" (U.S. Bureau of the Census 1910, 32).

The Bureau further instructed its operatives to count women "who regularly do outdoor farm work, even if on the family farm and not for wages" and to record as working any woman who "regularly earns money by some other occupation" in addition to her family's housework. The expansive instructions about women's work yielded much higher female gainful employment rates in 1910 than for surrounding census years, especially for married women (see Figure 3). The increase was particularly pronounced for black women, but the employment rate for white married women more than doubled as well, albeit from a much smaller base. The spike in married women's employment was especially pronounced in the South.

The Bureau backtracked in 1920, removing the opening statement emphasizing the importance of ascertaining the occupational status of all women and children. In farm work, which was seen as the main culprit in the "overcount," women in 1910 were to be reported with an occupation if they worked "regularly" outdoors on the farm. In contrast, in 1920 women had to work "regularly *and most of the time*," with work for "only a short time each day" not qualifying

as employment. The 1920 level of female employment, as expressed in census statistics, fell back in line with the previous historical trend, and 1910 would thereafter stand out as an anomaly in the series. Because these instructions dictated how the data were recorded on the census forms, the modern microdata derived from those forms preserve this temporary employment surge.

Figure 3: Married Female Labor Force Participation by Race, 1880-1990



Source: 1890 rates from C. Goldin. 1990. All others derived from S. Ruggles et al. 2024.

The 1910 occupation guidance is an extreme example of the impact of changes to census instructions, but most census questions underwent some evolution in this period. It behooves modern day historical researchers who use census microdata to be cognizant of the issues behind contemporary enumeration and data processing practices. Enumeration procedures and training

changed over time. Layers of field staff and data quality checks grew more complex with each decennial census. Preconceived notions informed by the social, cultural, and economic biases of the time informed "appropriate" census data on variables like occupation. The creation of a permanent Census Bureau provided leverage for its staff to establish institutional standards for training and data quality checks. Understanding how the data were created in the first place thus supports an informed use of the full count microdata in the present.

Microfilm

To address issues of storage, preservation, and accessibility, the Census Bureau undertook a "massive operation" to microfilm their collection of manuscript census forms (Genadek and Alexander 2022, 59). The pre-1950 census forms were microfilmed between 1937 and 1944, and the original paper versions of the manuscripts were destroyed.[2] Microfilm is now the primary source for the historical censuses, and any errors introduced during the microfilming process are irreversible. Such errors include missing pages, forms out of order, and poor image quality (U.S. Bureau of the Census 1966, 70). Most of the problems appear to be caused by human error rather than a failure of the technology. Although imperfect, the microfilming step was generally performed well. The issues tend to be small-scale and largely random, injecting a miniscule degree of noise that is unlikely to have a substantive impact on typical analyses.

Digital Data Capture

Genealogical companies became involved in large-scale census data transcription to enable their clientele to trace their family lineages. IPUMS subsequently reached agreements with Ancestry.com and FamilySearch to obtain their transcribed data for social science research. Ancestry primarily worked with contractors from East Asia to enter the data, while FamilySearch

[2] The 1890 census manuscripts had previously been destroyed by fire and are permanently lost.

worked with native-English-speaking volunteers. IPUMS had no involvement with this data entry for the 1900 to 1930 full count censuses, and there does appear to be some negative effects on the accurate transcription of person names and relatively verbose question fields when non-English speakers were used.

For the early twentieth-century full count databases, Ancestry performed their own data entry, and FamilySearch used this copy for their "verification" of data entry accuracy. The census files IPUMS received from Ancestry for 1900-1930 contained merged data from both genealogical companies. In many cases, there were multiple versions of a field, and the responses sometimes differed in terms of spelling and even substantively (e.g., Ancestry entered race as "white" and FamilySearch entered race as "black"). These competing fields required exploration and had implications for the data dictionaries used during IPUMS processing, as described below.

IPUMS received the data in two phases for each census. Phase I contained basic demographic fields (from Ancestry and FamilySearch) that the genealogy companies required for their business purposes. Phase II contained most of the remaining fields, which IPUMS contracted with Ancestry to enter (Ruggles 2023). This necessity to combine information from across files sometimes led to data merging issues. The additional variables for 1910 were sometimes shifted one or more rows within census pages, leading to cases of "infant farmers" or native-born persons with immigration information. We suspect the cause was a combination of data transcription error, data sorting error, and merging on the Ancestry side. We used logical checks (e.g., Household head has no occupation, but the spouse does, youngest/last person in the household has an occupation but the head does not, a man has fertility information, etc.) to identify pages with shifted records. We then corrected these cases with programming. The record shifts were not always consistent and sometimes led to missing information, which would ultimately be imputed

during the final stages of processing. We also discovered records out of order in 1920 and 1930, but to a lesser degree than in 1910, and those issues were more easily corrected.

A unique limitation to the 1910 full count data relates to group quarters. Unfortunately, Ancestry did not transcribe the institution field, inhibiting analysis of collective living arrangements. The issue impedes correctly specifying the universe of regular households and has implications for the construction of family interrelationship variables. IPUMS applied several fixes, including using the relationship-to-head field to infer group quarters (e.g., looking for "inmates"), identifying large households with more than five unrelated individuals, and using matched sample data to identify some group quarters (as explained below). But some group quarters are surely unidentified, and we cannot characterize institutions by type for the 1910 full count data.

For the 20th century data, the genealogy companies scanned the census microfilm as digital images from which they performed data entry. This shift to digital images undoubtedly increased productivity and made it financially feasible for IPUMS to contract for the additional data entry needed for the Phase II variables. Importantly, we have found no substantial negative effects resulting specifically from the microfilm imaging process.

The census image files were also highly beneficial for IPUMS processing. One of the most useful variables Ancestry provided in the census microdata was the image filename corresponding to each record. With this information, IPUMS staff could easily look up scanned images on the Ancestry web interface in real time to investigate data quality issues. Such work proceeded far more quickly than loading a microfilm reel and navigating to the specific page. Given the scale of the project, without this ready access to the scanned images, most manual data quality checks performed during IPUMS processing would have been impossible.

IPUMS Data Processing

The data provided by Ancestry.com are primarily literal transcriptions of the character string entries on the census forms. These string data contain spelling variations, errors, and other artifacts of the data capture process that make the original databases unsuitable for scientific research. IPUMS must clean the data and apply numeric classifications to categorical variables to make those variables usable. To avoid undercutting the business model of the collaborating genealogy company, the original strings themselves, including respondent names, are available only to researchers who apply for access to the restricted version of the data.

Initial Ancestry Data Assessment

Upon receiving the data, IPUMS assessed the data's completeness and general soundness. We first confirmed that the overall record count at the state level appeared plausible. The number of cases in the microdata should typically be slightly larger than the published contemporary counts. The general workflow thereafter included identifying duplicate and blank records, processing geography, creating household breaks, coding the data, and identifying other field-specific data quality issues. The process can be iterative. For example, while identifying duplicate records is one of the first steps performed on the data, we return to it once the detailed geography has been coded and we can compare the full count data to the published counts at finer levels.

Removing duplicate and blank records was the first step in cleaning the source data. "Blank records" are rows in the data that are not actual persons but were captured during data entry or errantly created during processing by Ancestry. Often, these blank records are informational lines inserted by the enumerator. We searched for records that were missing names or demographic information (age, sex, race, marital status, birthplace), and in some cases we performed manual reviews while validating geography. The typical rule to denote a false record required at least four

of these key variables to be blank. We also compared names between images within the same enumeration district to identify duplicate records. The 1910 data required additional work pertaining to the fraudulent records noted above. Using a combination of programming and manual review, we identified and removed fraudulent records by comparing sub-county population counts to the published data, as well as referring to known reports of fraud (U.S. Bureau of the Census 1911, 26).

The final step at this stage focused on identifying void pages where enumerators entered data, but census clerks determined the records either did not belong in that enumeration district or were enumerated elsewhere. In these cases, the whole census page was crossed out with the word "void" written over it. Ancestry had no mechanism for capturing such page-level edits. They entered these records and—given that they contained full demographic information—very few were removed during the blank record identification process. Moreover, the entries on these voided pages were not picked up as duplicates unless they were within the same enumeration district as their valid counterparts. Typically, we relied on geographic overcounts and manual inspection to identify these voided pages.

Geography and Household Breaks

Many verification checks require reference to geography, but this was complicated by ambiguities introduced by Ancestry's practice of only providing one geographic identifier for place of residence in their data entry. For example, one cannot separate an incorporated city from its surrounding rural township that shares the same name, which is not an uncommon situation. To address this, we used published figures to specify towns and places and manually reviewed the geography written on the original images. We updated the geographic information based on published names, population counts, and other documentation as needed. This process parsed the

geography into two fields: townships and places (villages, towns, cities). Not all states use the same geographic terminology for townships (e.g., beats, districts, wards), but all are functionally equivalent sub-county administrative units. Places include all incorporated settlements.

The next step required comparing the record counts at the county level to the published data collated by Michael Haines (Haines 2010). This process helped identify areas with over/undercounts relative to the published figures. Overcounts were typically duplicate entries stemming from data capture and, in some cases, incorrect county strings that produced errant county codes. County errors were often identifiable because another county in the same state had a surplus or deficit of similar magnitude. In some cases, we identified undercounts that spurred Ancestry to supply updated data. In other cases, the images of the forms are missing entirely and irretrievable, because the paper manuscripts were lost or the microfilming was flawed. Since only the microfilm now exist, there is no way to distinguish between these two sources of data loss.

One of the most important steps in IPUMS data processing involves distinguishing household groups within a database that arrives as a massive listing of individuals. We developed a variety of rules to denote household breaks, based on the available variables and particular features of each enumeration. First we needed to ensure that the sort order of the data was correct. All Ancestry data had a variable identifying the y-coordinate (vertical location) of the case on the census page, but this value was sometimes corrupted. The complementary line number variable appeared to be auto generated and was not always correct. We chose to prioritize the y-coordinate to sort records within a page, substituting line number when the y-coordinate was clearly wrong. Additional rules for each year focused on identifying characteristics of the first person of the household. These indicators include dwelling number sequence changes, surname changes from the previous household, a relationship-to-head response of "Head," and the presence of household-

level variables (e.g., farm status, home ownership, rent/value of home) on the person's record. None of these features were used independently to mark a household break but employing them together and in different combinations supported satisfactory household identification for most cases. Additional data quality checks included identifying households with multiple heads/spouses or out-of-order relationships (such as listing a spouse first and the head second) and flagging large households whose dissimilar surnames or lack of relationship variation suggested group quarters.

Data Dictionaries

The original string data are ultimately encoded into numeric classifications using dictionaries. The dictionaries are essentially correspondence tables that list all the string permutations that exist in the database, record their frequency, and provide a column to assign each string a numeric code in the output file. The benefit of this approach is its efficiency. One need only code a string once, regardless of how often it occurs in the data, and the dictionary can be sorted different ways and analyzed. It is also often possible to use the dictionary for one census as the starting point for another, typically leading to a large percentage of the cases being coded from the outset. There are, however, some liabilities to the dictionary approach to coding.

One of the drawbacks to dictionary encoding centers on the standardization by Ancestry of some of the responses during data transcription. Standardization reduces the number of strings, but when a problem arises, fixing the data requires manual inspection to disambiguate the responses. One example involves Canadian birthplaces. Enumerators were instructed to identify whether a respondent was born in English Canada or French Canada, and that was usually recorded on the forms (Department of the Interior 1900b; U.S. Bureau of the Census 1910; U.S. Bureau of the Census 1919a). However, many responses were simply standardized to "Canada" during data entry. Recovering the information would require a data entry operator to go through all Canada

strings in the data and manually update the strings, which was not feasible, so that original distinction between French and English Canadians was often not retained in the microdata.

A second, larger issue with the use of dictionaries concerns lack of context. Each string is coded in isolation from geography, household composition, or even the other characteristics of the person. All the records in which a string occurs receive the same code, regardless of any surrounding information. In 1900, the birthplace strings for Washington state and Washington D.C. were simply entered as Washington. While there were ways to correct some of these cases logically (e.g., many people who reported Washington state were living in Washington State and vice versa for Washington D.C.), the dictionaries cannot provide that broader context. At the individual level, an occupation response of "college" means something different for a 20-year-old probable student than a 60-year-old probable employee. It is simply not practical to add all the potentially relevant information into the dictionaries, which would grow exponentially with the addition of each new field and quickly become unmanageable. This anticipates the next challenge posed by the dictionary approach.

The sheer scale of some dictionaries can be problematic. For example, the 1900 occupation dictionary contains 2.8 million unique occupation strings, of which 1.5 million occur only once in the dataset. Coding these cases manually is impossible. There are methods to work around this issue using machine learning, probabilistic assignment, and string comparisons to identify similar cases—all of which were employed to differing degrees—but in the end, many strings in the large dictionaries remain uncoded. This does not present a significant problem for most analyses, because, while the number of uncoded strings may be high, they represent relatively few cases. Table 1 shows the proportion of uncoded cases and strings for each census year. It demonstrates that uncoded strings typically comprise about 2% of cases for the occupation and industry

variables. Moreover, a large percentage of these cases within the tail of the response distribution represent data quality issues with either the original enumerator entry or the data transcription. Such instances are best left to logical edits and imputation routines, described below.

Table 1: Occupation and Industry Classification Summary Statistics, 1900-1930

	Occupation 1950 Classification		Industrial 1950 Classification	
	Cases Uncoded	Strings Uncoded	Cases Uncoded	Strings Uncoded
1900[1]	1.9%	56.8%	3.4%	63.4%
1910	2.3%	17.2%	2.2%	12.9%
1920	1.5%	16.3%	1.4%	12.4%
1930	1.6%	15.1%	1.4%	11.3%

[1] 1900 only collected occupation; industry is inferred.

Source: S. Ruggles et al. 2024.

The final issue pertaining to data dictionaries stems from the explicitly intersecting nature of certain variables. The first type of situation concerns where the response in one variable affects the interpretation of another. For example, beginning in 1910 the census recorded occupation, industry, and class of worker (whether the respondent was an employer, wage employee, etc.). To facilitate analyses across census years, we consistently applied the occupation and industry categories and codes from the 1950 census. Coding occupation into the 1950 classification requires consideration of both industry and class of worker responses, as well as the occupation field itself. An example of this involves the occupation "farm worker," where a class of worker response of "self-employed" is a farmer, but a "wage employee" is a farm laborer—a distinction with significant socioeconomic implications. In the dictionary context, IPUMS incorporates all three work variables for coding purposes, but this makes the dictionary several times as large due to the explosion of combinations, harking back to the scale issue noted above.

A second type of variable intersection stems from the particulars of data entry for these censuses. The original data are the product of merging the indices between Ancestry and

FamilySearch, and two or more versions of many variables were included. This led to situations where the alternative string responses sometimes disagreed. Because of the number of cases and variables so affected, manual review was only possible for targeted diagnostic purposes. Our solution was to code each field independently and compare the results in what we term a "combo dictionary." This process served to identify substantive differences as opposed to non-meaningful spelling variations. If both fields received valid, differing codes, the final output value was assigned as "missing/not classified" and later allocated during the harmonization process (see below).

Despite these liabilities, dictionaries are the most feasible approach to the coding challenge posed by data of this scale. Hand-coding each record would take vastly greater resources than are available for this work, and hand-coding would pose its own set of problems, including inconsistency and resistance to systematic revision. On both those scores, the dictionaries fare well. The most practical alternative to dictionaries is an artificial intelligence approach, but that would require extraordinary care to develop and would pose its own challenges with respect to complexity and impenetrability. In sum, we determined that the best approach to coding strings at this scale was to treat them in isolation and try to correct obvious issues with selective logical edits and allocation.

Harmonization

After the strings are encoded, the data enters the final stage of processing: harmonization. Our aim was to make the data fully consistent with the rest of the IPUMS data series spanning 1850 to the present, including documenting the data within the web dissemination system. For the most part, significant variable recoding is not necessary at this point, because the IPUMS classifications were used as the target codes in the dictionaries. Nevertheless, some code tweaking

is inevitably required, many constructed variables are created, and various quality checks are conducted.

Data allocation and logical editing are a significant element of the harmonization stage. For most variables, missing values are allocated, which entails substituting a valid response from a person who shares key characteristics with the person under evaluation. As this implies, allocation happens at the individual level, unlike the dictionary encoding. The allocation matrices must be customized to each sample and variable, and the output is carefully examined for plausibility. Care must be taken when a blank response for a variable, such as literacy, is implicitly meaningful as opposed to a non-response. The former must not be allocated. Certain logical edits are performed after allocation to clean up the universe of respondents (e.g., men with children ever born) and correct inconsistencies. A few other selective edits are conducted, such as coding laborers to farm laborers if they resided on a farm. Table 2 shows the percentage of cases allocated or edited for selected variables. As is evident, the rates for a variable can differ substantially across samples.

Table 2: Percent of Cases Edited or Allocated by Variable, 1900-1930

	1900	1910	1920	1930
Age	1.5%	2.5%	3.5%	1.5%
Sex	0.4%	0.6%	0.6%	0.0%
Relationship	1.5%	2.2%	1.4%	1.2%
Marital status	0.8%	1.9%	0.6%	0.2%
Race	0.4%	0.5%	0.2%	0.1%
Literacy	3.5%	2.5%	1.3%	0.8%
School attendance	--	4.4%	0.0%	0.2%

Source: S. Ruggles et al. 2024.

Data Evaluation

We implemented a number of quality checks to verify that the microdata accurately represent the historical population as transcribed in the Census. One set of checks involved comparisons to the published census volumes. For most variables these comparisons involved national-level overviews of results and ad hoc comparisons conducted during the exploration of suspected data irregularities. Geography, however, was scrutinized systematically at the county and even the sub-county level. While we do not expect the number of cases to match exactly between the microdata and the published census tables, we do expect them to be close, unless there is a known issue regarding coverage. The top portion of Table 3 indicates the number of counties under- and over-counting the population by more than one percent and five percent both before and after data processing by IPUMS. As noted previously, over-counting is generally more common because of duplicates and other invalid records. As the table indicates, the number of errant counties declined significantly after IPUMS processing that aimed to eliminate such cases. In the initial 1900 data delivery from Ancestry, 922 counties in the microdata had an overcount greater than 1% relative to the published numbers, and 136 counties had an overcount greater than 5%. IPUMS processing reduced the overcount to 194 and 21 counties respectively, and many of these are small counties with high variability. The quality of the data clearly improves over time, and the net overcount in all years is under 0.1% of the national population. In short, the data generally represent the population totals accurately and can serve as a source for calculating population denominators for other data sources (Antoine-Jones et al. 2023, 495; Eiermann et al. 2022, 1961).

Table 3: Microdata Reporting Rates Compared to Published Counts, 1900-1930

	Error Rate	Data Stage	1900	1910	1920	1930
Number of Overreporting Counties	> 1%	Original Data Delivery	922	128	67	42
		Post IPUMS Processing	194	81	46	15
	> 5%	Original Data Delivery	136	21	8	3
		Post IPUMS Processing	21	8	3	1
Number of Underreporting Counties	> 1%	Original Data Delivery	26	17	20	18
	> 5%	Original Data Delivery	17	5	5	0
Population Overreported after Processing			270,236	100,753	97,934	65,464
Population Underreported after Processing			201,083	29,401	78,602	62,996
Total Population Overreporting			69,153	71,352	19,332	2,468
Total Population Overreporting Rate			0.09%	0.08%	0.02%	0.02%

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

The second major data quality check involved matching the full count data to the sample data for each census at the individual record level (Ruggles et al. 2023). In the years prior to development of the full count databases, IPUMS completed a series of projects to create nationally representative (1 to 5%) microdata samples of each of the censuses from 1900 to 1930. The sample data were entered directly from microfilm by data entry operators at the University of Minnesota, with research staff conducting data quality checks. A fraction of cases went through a blind verification stage in which data were entered a second time and compared to the original data entry.

In matching the full count data to the samples, we calculated a disagreement rate between the two sources for each variable. We term this a disagreement rate because we do not know which value is correct, or if both are incorrect. But our default assumption is that the sample data are generally of higher quality. In creating the sample data, we had much more labor per case to devote to data cleaning; and the samples were developed specifically with research in mind, with the entire operation overseen by social scientists, rather than originating as a genealogical product. The disagreement measure helped identify bad codes in the dictionaries but also potential systematic full count data transcription problems. Table 4 shows the agreement rates by variable and census

year for select variables for 1900-1930.[3] These rates are calculated before any logical editing and allocation is performed on the data and are limited to cases within the logical universe of respondents (i.e., those who should have answered the question) for each variable from the samples.

Table 4: Agreement Rates between 1900-1930 Full Count Data and Samples

Variable Name	Variable Description	1900	1910	1920	1930
Household					
OWNERSHP	Home Ownership	92.6%	97.1%	98.6%	98.7%
FARM	Farm Status	94.4%	97.7%	98.1%	99.4%
Demographic					
RELATE	Relationship to Head Status	98.0%	98.5%	98.1%	99.1%
SEX	Sex	99.7%	99.0%	99.3%	99.5%
AGE	Age	94.3%	95.8%	97.7%	97.0%
MARST	Marital Status	99.0%	99.0%	98.4%	99.2%
Race, Ethnicity, and Nativity					
RACE	Race	99.4%	99.3%	99.4%	99.7%
BPL	Birthplace	98.0%	96.9%	95.5%	99.1%
FBPL	Father's Birthplace	97.1%	97.5%	96.4%	98.7%
MBPL	Mother's Birthplace	97.2%	97.3%	95.9%	98.6%
YRIMMIG	Year of Immigration	91.9%	90.3%	91.7%	95.4%
Education and Work					
SCHOOL	School Attendance	95.3%	96.7%	98.8%	98.5%
LIT	Literacy	95.0%	98.2%	98.5%	98.9%
OCC1950	Occupation 1950 Classification	92.9%	94.7%	89.2%	92.5%
IND1950	Industry 1950 Classification	93.3%	90.1%	89.9%	91.1%
CLASSWKR	Class of Worker	n.a.	97.2%	96.6%	97.7%

"n.a." indicates the variable is not available in that census.
Source: Ruggles et al. 2023, 2024.

For most variables, the full count and sample data encoding agree roughly 95 to 97% of the time. Variables commonly used for record linkage such as sex, age, and race display high agreements rates in all years. Fields with more potential entries, such as home value or rent, and complex variables with many categories, such as occupation, exhibited lower agreement rates.

[3] For the full list of variables, universes, and notes on comparisons, please consult table A1 in the Appendix.

Rates of agreement also tend to be lower when the full count data offer multiple variants of the transcribed field. If the alternative transcriptions yielded valid but conflicting results, the output was coded as missing and imputed; but in 78% of those cases, one of the competing entries in the full count data matched the code in the sample. Thus, the agreement rates for some variables can exaggerate the degree of difference between the sources, and the quality of the full count data are higher than these figures suggest.

Some of the disagreements between the sample codes and full-count data codes appear to stem from IPUMS processing. When looking at occupation in 1930, we find some results that suggest a combination of data entry errors and dictionary coding errors. For cases with occupation codes that disagreed in 1930, 26% had different codes, and the Jaro-Winkler similarity score was less than 0.8, suggesting very different data transcriptions. (Jaro-Winkler similarity scores range in value from 0 (least similar) to 1 (most similar)). However, we find 9% of cases had a Jaro-Winkler similarity score between 0.8 and 0.9, 31% between 0.9 and 1.0, and almost 34% of the cases had the exact same string. This suggests that in some cases we coded strings incorrectly in one of the databases—most likely the full count version. The difficulties of scale and multiple interacting variables made occupation coding challenging, and when it was conducted by different teams using a variety of strategies, the result sometimes differed.

Overall, the IPUMS full count data match the published population totals and the sample data well. Data transcription errors and IPUMS processing decisions explain most discrepancies between the full count data and the samples. Importantly, variables typically used for record linkage have high agreement rates, allowing researchers to link external historical datasets to the full count census data with confidence. Most records are not logically edited or imputed, although researchers can download the data quality flags for variables to investigate or remove cases as

needed. For researchers concerned about potential errors who are conducting national-level analyses, validating results using the sample data can help assess whether errors of this nature are influencing their analyses.

Conclusion

The early 20th-century U.S. Census microdata underwent a long journey from several hundred million responses communicated face-to-face to enumerators to a digital format suitable for computerized data analysis. These records were not only collected within a particular historical context, but preservation and harmonization decisions also impacted the final data. This paper summarizes the historical context and decisions made within the three phases of data production: historical census enumeration, digital data capture, and data processing. While some of these issues represent unique historical artifacts or errors that were corrected for general data analysis, decisions within each of these stages of data production and processing have implications for research. Users of these rich data need to consider how this production history might affect the data's suitability for particular research questions.

The full count microdata sets are very large, complicating and constraining the options for processing. If users are performing national-level analysis, they may wish to perform parallel analyses using the sample data, which we presume to be of generally superior quality because the samples were designed with social science research in mind. A positive result from that comparison should give a researcher confidence that they can exploit the greater scale of the full count data without being overly concerned about the potential effect of transcription issues. Despite their limitations, the IPUMS full count data represent a powerful resource to study small sub-populations, perform contextual and local analyses at fine geographic detail, and link external

resources to the IPUMS data series. The sheer scale of this data offers the prospect to answer entirely new economic and demographic questions.

Acknowledgments

This project was supported in part by funding from the Minnesota Population Center (P2CHD041023) and by a grant from the Eunice Kennedy Shriver National Institute for Child Health and Human Development (R01-HD078322). Many individuals at the Minnesota Population Center contributed to the project, but we wish to especially acknowledge the work of Ronald Goeken, who contributed significantly to all aspects of the project over a five-year period. We are grateful to Miriam King for her helpful suggestions.

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

"Twelfth Census of the United States, Enumerator's Street Book," NARA, Instructions File #703.

"Thirteenth Census of the United States, 1910," (NARA microfilm publication T624, 1,178 rolls). Records of the Bureau of the Census, Record Group 29. National Archives, Washington, D.C.

"Thirteenth Census of the United States, Enumerator's Street Book," NARA, Instructions File #128.

"Qualifications, Duties, and Compensation of Census Enumerators," NARA RG 29, 215/232: F-4. Supervisors received numerous form letters (fifty in all) and other miscellaneous correspondence. See NRA RG 29,212/181: 25.

Untitled press release, 29 November [1919], NARA, RG 29, 198/164: 51.

"Instructions For Clerks Examining Population Schedules to Take Effect June 11," NARA RG 29, Memo and Notes of J.A. Hill, File P-4 Population Instructions 1910 Census. "Instructions for Editing: Family Card, form 8-3315" NARA RG 29, Instructions File No. 6.

"List of Cards to be Rejected in the Sort of the Fourteenth Census Occupation Cards for Females and Males," NARA RG 29.

An Act to provide for the Thirteenth and subsequent decennial censuses. 1909. H.R. 1033, July 2, 1909.

Ancestry.com. *1910 United States Federal Census* [database on-line]. Lehi, UT, USA: Ancestry.com Operations Inc, 2006.
https://www.ancestryinstitution.com/imageviewer/collections/7884/images/4454885_01028

Anderson, M. 2010. The Census and the Federal System: Historical Perspectives. *The Annals of the American Academy of Political and Social Science* 631(1):152-162.
https://doi.org/10.1177/00027162103737

Anderson, M.J. 2015. "Building the Federal Statistical System in the Early Twentieth Century." Chapter in *The American Census: A Social History.* Yale University Press.

Annual Report of the Director of the Twelfth Census. 1900. 56(2), House Doc. No. 5, November 1, 1900.

Antoine-Jones, A., J. J. Feigenbaum, L. Hoehn-Velasco, C. Muller, and E. Wrigley-Field. 2023. Racial Inequality in the Prime of Life: Infectious Disease Mortality in U.S. Cities, 1906-1933. *Social Science History* 47:491-504.

Conk, M.A. 1981. Accuracy, Efficiency and Bias: The Interpretation of Women's Work in the U.S. Census of Occupations, 1890-1940. *Historical Methods* 14:65-72.

Cummings, J. 1913. The Permanent Census Bureau: A Decade of Work. *Publications of the American Statistical Association* 13 (104): 605-638.

DeBow, J.D.B. 1853. *Compendium of the Seventh Census*. Washington: Robert Armstrong.

Department of Interior, Census Office. 1900a. *Census Act of March 3, 1899*. Washington: GPO.

Department of the Interior, Census Office. 1900b. *Instructions to Enumerators*. Washington: GPO.

Durand, E.D. 1910. Changes in the Census Methods for the Census of 1910. *American Journal of Sociology* 15(5):619-632.

Durand, E.D. 1913. The Census Methods of the Future. *Journal of the American Statistical Association* 13(104):563-582.

Eckler, A.R. 1972. *The Bureau of the Census.* Praeger Publishers.

Eiermann, M., E. Wrigley-Field, J. J. Feigenbaum, J. Helgertz, E. Hernandez, and C. E. Boen. 2022. Racial Disparities in Mortality During the 1918 Influenza Pandemic in United States Cities. *Demography* 59(5): 1953-1979.

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

Goldin, C. 1990. *Understanding the Gender Gap: An Economic History of American Women*. New York: Oxford University Press.

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

Holt, W.S. 1929. *The Bureau of the Census: Its History, Activities and Organization*. Washington: The Brookings Institution.

Instructions for Editing: Family Card, form 8-3315. NARA RG 29, Instructions File No. 6.

Instructions For Clerks Examining Population Schedules to Take Effect June 11. NARA RG 29, Memo and Notes of J.A. Hill, File P-4 Population Instructions 1910 Census.

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. 2012. Decennial Censuses: 1920 Census in *Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey*. Anderson, M.J., Constance F. Citro, and Joseph J. Salvo, editors. CQ Press, 2012 2e.

Merriam, W.R. 1901. *Report of Director of the Twelfth Census to the Secretary of the Interior, June 30, 1901*. Washington: GPO.

"Permanent Census Office." 1901. *Publications of the American Statistical Association*, Volume 7, Issue 56.

Qualifications, Duties, and Compensation of Census Enumerators. NARA RG29, 215/232: F-4.

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. 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., 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

Thirteenth Census of the United States, 1910 (NARA microfilm publication T624, 1,178 rolls). Records of the Bureau of the Census, Record Group 29. National Archives, Washington, D.C.

Truesdell, L.E. 1965. *The Development of Punch Card Tabulation in the Bureau of the Census, 1890-1940 With Outlines of Actual Tabulation Programs*. Washington: GPO.

Untitled press release. 29 November [1919], NARA, RG 29, 198/164: 51.

U.S. Bureau of the Census. 1907. *Report of the Director to the Secretary of Commerce and Labor, Concerning the Operations of the Bureau for the Year 1906-07*. Washington: GPO.

U.S. Bureau of the Census. 1908a. *Report of the Director to the Secretary of Commerce and Labor, Concerning the Operations of the Bureau for the Year 1907-08*. Washington: GPO.

U.S. Bureau of the Census. 1908b. *Report of the Director to the* Secretary *of Commerce and Labor: Concerning the Operations of the Bureau for the Year 1908-09*. Washington: GPO.

U.S. Bureau of the Census. 1910. *Instructions to Enumerators*. Washington: GPO.

U.S. Bureau of the Census. 1911. *Report of the Director to the Secretary of Commerce and Labor: Concerning Operations of the Bureau for the Year 1909-10*. Washington: GPO.

U.S. Bureau of the Census. 1913. *Annual Report of the Director of the Census to the Secretary of Commerce and Labor for the Fiscal Year Ended June 30, 1912.* Washington: GPO.

U.S. Bureau of the Census. 1919a. *Instructions to Enumerators*. Washington: GPO, 1919.

U.S. Bureau of the Census. 1919b. *Instructions to Supervisors of Census, General Instructions.* Washington: GPO.

U.S. Bureau of the Census. 1919c. *Instructions to Supervisors of Census, Oral Instruction of Enumerators and Conduct of the Enumeration.* Washington: GPO. NARA RG 29, 212/181: 25.

U.S. Bureau of the Census. 1919d. *Instructions to Supervisors of Census, Employment of Inspectors to Supervise the Work of Enumerators in Cities.* Washington: GPO. NARA RG 29, 212/181: 25.

U.S. Bureau of the Census. 1920a. *The Act for Providing for the Fourteenth Decennial Census, taken January 1, 1920*. Washington: GPO.

U.S. Bureau of the Census. 1920b. *Annual Report of the Director of the Census to the Secretary of Commerce, June 30, 1920*. Washington: GPO.
U.S. Bureau of the Census. 1930. *Instructions to Enumerators*. Washington: GPO, 1930.

U.S. Bureau of the Census. 1966. *1960 Population and Housing Censuses*. Washington: GPO.

U.S. Census Bureau. 2002. *Measuring America: The Decennial Censuses from 1790 to 2000*. Washington: GPO.
U.S. Census Bureau. "1910 Overview - History - U.S. Census Bureau." United States Census Bureau, 2008. https://www.census.gov/history/www/through_the_decades/overview/1910.html.
U.S. Census Bureau. "1920 Overview - History - U.S. Census Bureau." United States Census Bureau, 2008. https://www.census.gov/history/www/through_the_decades/overview/1920.html.

U.S. Census Office. 1900. *The Report of the Director of the Twelfth Census to the Secretary of the Interior for the Fiscal Year Ended June 30, 1900*. Washington: GPO.

U. S. Census Office. 1902. *Act of March 6, 1902, Providing for the Establishment of a Permanent Census Office.* Washington: GPO.

Willcox, W. F. 1914. The Development of the American Census Office Since 1890. *Political Science Quarterly* 29(3):438-459.

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

Appendix

Table A1: Agreement Rates between 1900-1930 Full Count Data and Samples

Variable Name	Variable Description	Universe	1900	1910	1920	1930
Household						
OWNERSHP	Home Ownership	Head	92.6%	97.1%	98.6%	98.7%
VALUEH	Value of House	Head, Own, Non-farm	n.a.	n.a.	n.a.	91.0%
RENT30	Rent	Head, Rent, Non-farm	n.a.	n.a.	n.a.	88.6%
RADIO30	Owns Radio	Head	n.a.	n.a.	n.a.	99.1%
FARM	Farm Status	Head	94.4%	97.7%	98.1%	99.4%
Demographic						
RELATE	Relationship to Head Status		98.0%	98.5%	98.1%	99.1%
SEX	Sex		99.7%	99.0%	99.3%	99.5%
AGE	Age		94.3%	95.8%	97.7%	97.0%
BIRTHYR	Year of Birth		96.1%	[4]	[4]	[4]
BIRTHMO	Month of Birth		99.3%	n.a.	n.a.	n.a.
MARST	Marital Status	Age 12+	99.0%	99.0%	98.4%	99.2%
AGEMARR	Age at First Marriage	Age 12+, Currently Married	n.a.	n.a.	n.a.	95.6%
DURMARR	Duration of Current Marriage	Age 12+, Currently Married	96.1%	96.3%	n.a.	n.a.
CHBORN	Children Ever Born	Age 12+, Ever Married[1]	94.8%	97.1%	n.a.	n.a.
CHSURV	Children Surviving	Age 12+, Ever Married[1]	94.9%	97.7%	n.a.	n.a.
Race, Ethnicity, and Nativity						
RACE	Race		99.4%	99.3%	99.4%	99.7%
BPL	Birthplace		98.0%	96.9%	95.5%	99.1%
FBPL	Father's Birthplace		97.1%	97.5%	96.4%	98.7%
MBPL	Mother's Birthplace		97.2%	97.3%	95.9%	98.6%
LANGUAGE	Language Spoken at Home	Age 10+	n.a.	97.4%	n.a.	n.a.
MTONGUE	Mother Tongue	Foreign-Born	n.a.	[5]	93.6%	98.1%
YRIMMIG	Year of Immigration	Foreign-Born	91.9%	90.3%	91.7%	95.4%
YRSUSA	Years in the United States	Foreign-Born	91.8%	[6]	[6]	[6]
CITIZEN	Citizenship Status	Foreign-Born[2]	84.9%	91.7%	96.4%	97.7%
SPEAKENG	Speaks English	Age 10+	96.1%	98.0%	98.3%	99.3%
Education and Work						
SCHOOL	School Attendance	[3]	95.3%	96.7%	98.8%	98.5%
LIT	Literacy	Age 10+	95.0%	98.2%	98.5%	98.9%
OCC1950	Occupation 1950 Classification		92.9%	94.7%	89.2%	92.5%
IND1950	Industry 1950 Classification		93.3%	90.1%	89.9%	91.1%
CLASSWKR	Class of Worker		n.a.	97.2%	96.6%	97.7%
N			*755,679*	*918,047*	*1,050,634*	*6,095,331*
Linkage Rate			*96.4%*	*95.7%*	*98.7%*	*98.6%*

Source: Ruggles et al. *IPUMS USA: Version 14.0* [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D010.V14.0; Ruggles et al. *IPUMS Ancestry Full Count Data: Version 4.0* [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D014.V4.0

"n.a." indicates the variable is not available in that census.

[1] 1900 Universe includes all females age 12+, in 1910 the universe includes all ever-married females age 12+.

[2] 1900-1910 Universe includes males age 21+ who were not citizens at birth. 1920-1930 universe includes all persons who were not citizens at birth.

[3] School Universe varied between censuses, but generally anybody could respond with a valid school attendance so there is no universe restriction for this analysis.

[4] Birth year is constructed from age 1910-1930 but was a separate question in 1900.

[5] 1910 Mother Tongue was not transcribed in the Complete Count Data.

[6] Years in the United States is constructed for 1910-1930 but was a separate question in 1900.