Improving Historical Census Transcriptions: A Machine Learning Approach*

Christian Møller Dahl[a], Sam Il Myoung Hwang[b],
Torben S.D. Johansen[a], and Munir Squires[b]

[a]University of Southern Denmark
[b]University of British Columbia

November 21, 2024

Abstract

Historical census records enable researchers to track individual outcomes over time, but linking individuals across census rounds is particularly challenging for minority and immigrant populations due to transcription errors in handwritten names. We develop a machine learning approach that improves name transcription in historical U.S. census records, addressing the specific challenges of transcribing unfamiliar names and dense tabular formats. Independent human transcribers disagree on names in 30 percent of records, with higher disagreement rates for foreign-born individuals and non-English speakers. Our machine transcriptions increase linking rates by 147 percent for records where human transcribers disagree, while simultaneously improving match quality by 38 percent. These improvements help expand sample sizes for traditionally under-linked groups - including the foreign-born, non-white residents, and those with no formal schooling - where each additional linked record is particularly valuable for statistical inference. Validation against independent genealogical records confirms these gains represent genuine accuracy improvements rather than spurious matches. Our findings demonstrate that improved transcription methods can substantially enhance research on historically underrepresented populations in linked census data.

Keywords: Census linkage, machine learning transcription, historical microdata
JEL Classification: C55, C81, N00

*This research was undertaken thanks to funding from the Canada Excellence Research Chairs program awarded to Dr. Erik Snowberg in Data-Intensive Methods in Economics. We thank the participants of the Canadian Economic History Network sessions at the 2024 Canadian Economic Association Conference for their helpful comments. Correspondence can be addressed to ilmyounghwang@gmail.com.

1 Introduction

Tracking individual outcomes across time is central to many fundamental questions in economics, from understanding intergenerational mobility to measuring the long-run impacts of historical events. The U.S. historical census datasets from 1850 to 1950 offer unprecedented potential for such analyses, containing socio-demographic information for approximately 900 million person-years. However, a crucial challenge limits researchers' ability to exploit this rich data source, particularly for minority and immigrant populations: accurately transcribing handwritten names from historical census forms. The dense tabular format of census records, variation in handwriting styles across thousands of enumerators, and physical degradation of documents make accurate transcription particularly difficult. These challenges are especially acute for non-English names and those of minority populations, which were more likely to be misrecorded or anglicized. Because of these challenges, even the best current methods can link only about half of male records across adjacent census rounds, with substantially lower rates for non-white Americans and immigrants (Buckles et al., 2023).

In this paper, we develop a novel machine learning approach to improve handwriting recognition in historical census records, leading to substantial gains in census linking outcomes particularly for traditionally under-linked populations. Our method differs from existing approaches in three key ways. First, we utilize two independent human transcriptions of the 1940 census records to build reliable training data and identify records most in need of improvement, with particular attention to names from minority and immigrant populations. Second, we develop a specialized pipeline that processes census forms more effectively than general-purpose optical character recognition (OCR) systems, which typically struggle with handwritten data and dense tabular formats. Third, we process handwriting character-by-character rather than as complete words, allowing us to avoid biasing transcriptions toward common anglicized names and to provide uncertainty measures for downstream linking decisions.

Our analysis yields three main results. First, for records where human transcribers disagree—29.9 percent of our sample—our improved transcriptions increase the linking rate from 8.4 to 20.8 percent, representing a 147 percent improvement. This dramatic increase in linking success is accompanied by a 38 percent increase in match quality, as measured by the share of links validated through matching middle initials. When averaged across all records (including those where human transcribers agreed), our improvements increase the overall linkage rate by 12 percent.

Second, we find that our approach performs best precisely where it is most needed: the impact on linking rates is largest in enumeration districts with the lowest legibility, where linking rates increased by up to 35 percent. This pattern suggests our method is particularly valuable for addressing the most severe biases in linked samples.

Third, validation against genealogical records confirms that our improvements represent genuine accuracy gains rather than spurious matches. When compared to FamilySearch.org profile names, which incorporate multiple historical sources, our machine learning transcriptions achieve higher match rates than either Ancestry or FamilySearch

census transcriptions for records where human transcribers initially disagreed.

This paper advances two distinct literatures. First, we contribute to research on historical census linking and the construction of longitudinal microdata. This literature has primarily focused on developing better matching algorithms and expanding data sources. Early work by Ferrie (1996) established manual methods for linking individuals across censuses, setting standards for match quality that influenced subsequent approaches. Abramitzky et al. (2012) and Abramitzky et al. (2014) automated these processes, enabling large-scale linking projects. Recent advances have dramatically improved both the scale and representativeness of linked samples. The Census Tree Project leverages genealogical research and machine learning methods trained on family trees to create over 700 million census-to-census pairs, achieving the highest linkage rates to date (Buckles et al., 2023; Price et al., 2021).

However, this literature has largely treated name transcriptions as fixed inputs to the linking process. While careful attention has been paid to handling phonetic name variations and standardizing common nicknames, the quality of the underlying transcriptions themselves has been taken as given. Recent work by Ghosh et al. (2024) shows that transcription quality materially affects linking success, and that poor legibility creates substantial selection bias in linked samples. Our contribution is to demonstrate that machine learning can systematically improve transcription quality, particularly for traditionally under-linked populations. Unlike approaches that rely on specialized data sources like family trees (Hwang and Squires, 2024), our method can be applied using a generalizable approach that scales to the full census.

Our second contribution is to the literature on automated historical document processing. The dominant approach relies on general-purpose optical character recognition (OCR) methods, which struggle with handwritten data and tabular formats. While specialized frameworks like LayoutParser (Shen et al., 2021) have improved processing of historical documents such as newspaper archives (Dell et al., 2024), they face challenges with the dense tabular structure of census records. Our key methodological innovation is to break down census transcription into discrete steps—first identifying table structure, then extracting individual cells, and finally processing characters independently rather than as complete words. This approach yields three advantages: (1) it processes tabular data more effectively than general OCR systems, (2) it avoids biasing transcriptions toward common names by making character-level rather than word-level predictions, and (3) it provides uncertainty measures that can inform downstream linking decisions.

Our work demonstrates that substantial gains in historical data construction can come from addressing fundamental measurement issues rather than focusing solely on linking algorithms or additional data sources. By exploiting the unique characteristics of census records—standardized layouts, constrained character sets, and rich metadata—through specialized machine learning approaches, we achieve dramatic improvements in both the quantity and quality of linked records. Moreover, by focusing on transcription quality rather than linking algorithms, we show that substantial gains in historical data construction can come from addressing fundamental measurement issues. The scale of our

improvements—doubling linking rates for targeted records—suggests that transcription quality may be as important as matching methodology for improving linked historical data.

The remainder of this paper proceeds as follows. Section 2 describes our machine learning transcription system. Section 3 outlines our census data and linking framework. Section 4 demonstrates that our improved transcriptions dramatically increase linkage rates, with the largest gains occurring for low-legibility records that human transcribers find most challenging.

2 Machine Learning Transcription

We develop a specialized machine learning pipeline for transcribing handwritten names from census records. Our approach differs from general handwritten text recognition systems by focusing specifically on the challenges of census forms: dense tabular layouts, varying handwriting styles across enumerators, and the need to process millions of standardized records efficiently. Figure 1 illustrates these challenges with a representative census form from Rhode Island. The figure demonstrates the typical layout with its dense tabular structure, where names and other information must be precisely located and extracted. The cramped spacing between rows and columns further complicates accurate transcription, as letters sometimes overlap or extend beyond their intended cells. While a human reader can distinguish these overlapping characters through context, traditional OCR systems often fail at this task, highlighting the need for a specialized approach.

The consistency of census form layouts across millions of records allows us to optimize our system specifically for this format. This section outlines our core methodology for training data construction, model architecture, and performance metrics. Complete technical details are provided in Appendix A.

2.1 Training Data Construction

The quality of machine learning transcription depends critically on the training data, but ensuring accuracy in historical handwritten records poses significant challenges. Most historical datasets rely on a single human transcription, which is typically treated as ground truth despite potential errors. The U.S. census is particularly valuable for developing robust handwriting recognition systems due to its extensive coverage by genealogical organizations and the availability of dual transcriptions. From 1860 to 1940 (except 1880), each census round has been independently transcribed by both Ancestry.com and FamilySearch.org, providing a rich source of validated training data. The frequency of disagreement between transcribers—approximately 30 percent in our sample—highlights how common transcription errors are. This wealth of dual-transcribed historical records, driven by strong genealogical and research interest in U.S. censuses, provides an exceptional opportunity to develop and validate handwriting recognition methods that could be applied to other historical documents with limited or single transcriptions.

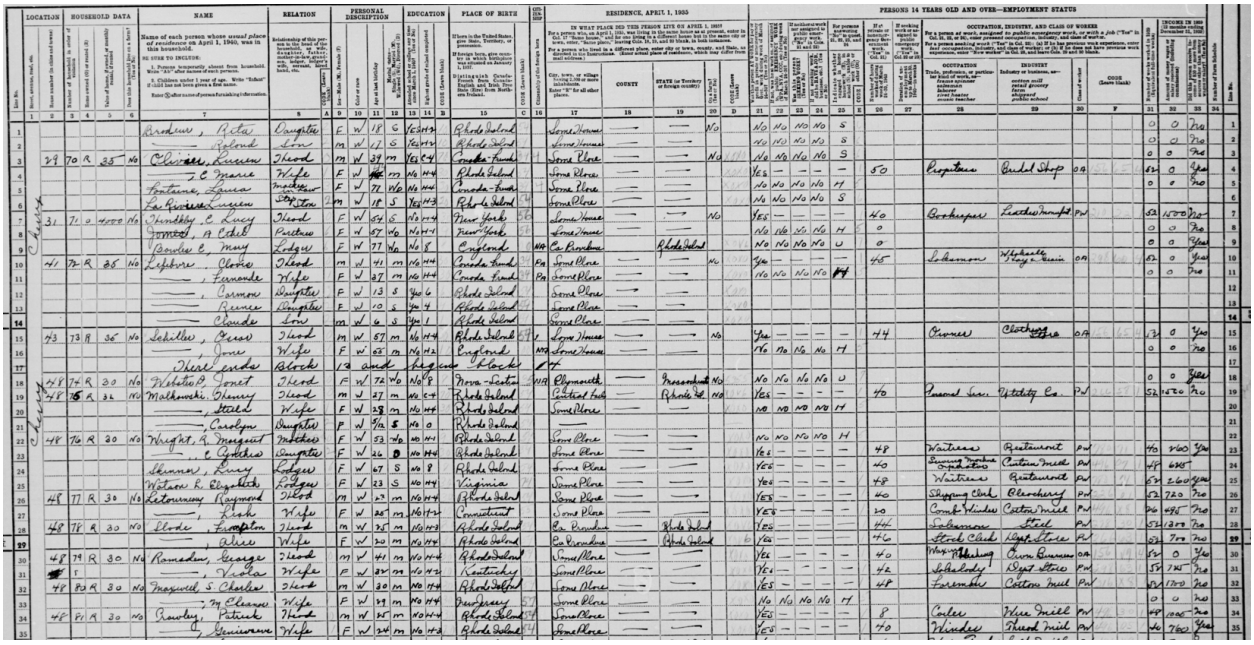


Figure 1: An example of a census form with average legibility

Note: This is an image of a form from the 1940 census for Enumeration District 4-252 of Rhode Island. This enumeration district was chosen because the legibility of its census forms is closest to the mean legibility for the state. We measure legibility by the percentage of records for which the transcriptions from Ancestry.com and FamilySearch.org agree. (Ghosh et al., 2024).

To ensure high-quality labels for training, we only include records where both transcriptions agree exactly on first name, last name, and age. From our initial sample of Rhode Island and selected regions of other states, this approach provides 392,085 training examples for first names and 121,465 for last names, numbers that will increase substantially as we expand to the full census.

There are substantially fewer last name training examples than first names due to our conservative handling of "ditto" marks in the census forms. Census enumerators frequently used ditto marks (") to indicate that an individual shared their last name with the person listed above them. Because our transcribed string data does not distinguish between actual written last names and ditto marks, we exclude all cases where consecutive records share the same last name. While this approach discards some valid training examples, it ensures our model learns from clear, unambiguous cases.

Our training data includes 29,641 unique first names and 28,032 unique last names. The similar counts suggest comparable diversity in first and last names, though first names include variations like "James F" versus "James W". We reserve 10% of records as a test set for measuring out-of-sample performance.

2.2 Model Architecture and Pipeline

Our transcription pipeline addresses two fundamental challenges: identifying where names are located on census forms, and reading the handwritten text in those locations. Both tasks have traditionally been performed by human transcribers, who intuitively understand how to scan a census page's structure and interpret handwriting. Our approach breaks these human capabilities into discrete computational steps.

The first challenge - finding names on the page - is surprisingly difficult for computers. While humans easily recognize the grid structure of census forms, computers need explicit guidance to understand this organization. Previous approaches like LayoutParser (Shen et al., 2021) were designed for newspaper articles, where text flows in columns. These methods struggle with census forms, which pack information densely into tables.

We developed a specialized approach for census tables that mirrors how a human might process the page structure. The process begins by identifying all horizontal and vertical lines that make up the table grid. Then, the system finds where these lines intersect to locate individual cells. Finally, it extracts the content of cells containing names. This systematic decomposition allows our system to reliably locate and isolate each piece of relevant information.

To handle cases where pages are warped or skewed from scanning - similar to how a human can read a slightly tilted page - we process each page in three ways: looking at the full page, just the top half, and just the bottom half. This redundancy helps minimize transcription errors when parts of a page are degraded or distorted.

Once names are located on the page, we turn to the second challenge: reading the handwritten text itself. Here, we diverge from traditional approaches in ways that may seem counterintuitive but which we have found prove more effective. Most handwriting

recognition systems try to process text as a sequence, similar to how humans read. Instead, we treat each character position independently, more like how a computer might process a digital form with separate boxes for each letter. This simplification works well for census names because they have relatively predictable lengths, each character typically occupies its own space, and the set of possible letters is known (a-z plus basic separators). This constrained structure makes independent character recognition particularly effective for census records.

Our character-by-character approach offers several key advantages over traditional word-based recognition methods. Consider a difficult-to-read surname where the first letter could be either 'S' or 'B'. A system trained to recognize complete words might lean toward transcribing it as "Smith" rather than "Brith" simply because "Smith" is a more common surname. Our approach avoids this bias by making each letter prediction independently. For each character position, the model outputs a probability distribution over possible letters. When transcribing an ambiguous first letter, the model might assign 60% probability to 'S', 35% probability to 'B', and 5% probability distributed across other letters.

This granular uncertainty quantification proves valuable in two ways. First, it helps prevent systematic biases toward common names that could distort linked samples. Second, it provides detailed confidence measures that can inform downstream linking decisions. A linking algorithm would ideally treat a name with high-confidence transcriptions differently from one where several characters have ambiguous readings.

To further improve accuracy, we train separate models for first and last names, allowing each model to specialize. The first name model learns common patterns in given names (like the frequency of "John" or "Mary"), while the last name model focuses on surname patterns. Due to the common use of ditto marks (") to indicate repeated surnames in census records, we have fewer training examples for last names than first names. To overcome this data limitation, we first train the model on the larger set of first names, then adapt this knowledge to the task of recognizing last names through transfer learning.

A crucial feature of our system is that it provides confidence scores for each character it transcribes. When a character is ambiguous - for instance, if it could be either an 'a' or an 'o' - the model expresses this uncertainty rather than making an arbitrary choice. These confidence scores prove valuable for downstream linking tasks, as they allow us to identify cases where transcription uncertainty might affect matching decisions.

Our approach represents a significant departure from existing census transcription methods. Traditional optical character recognition (OCR) systems expect clean, printed text and struggle with handwriting. More advanced handwriting recognition systems often try to process entire pages at once, which becomes computationally intensive and error-prone with dense census formats. By breaking the problem into smaller pieces - first finding name locations, then reading individual names - we achieve higher accuracy while keeping computational requirements manageable.

The tradeoff is that our system requires significant upfront investment in specialized models for census forms. However, once trained, the system processes new pages quickly

and consistently, with costs that scale roughly linearly with the number of records. This makes it particularly suitable for large-scale applications like full census transcription.

2.3 Model Performance

Table 1 presents our model's out-of-sample performance on the test set. We evaluate performance using two metrics: sequence accuracy and token accuracy. Sequence accuracy requires every character in a name to be correctly transcribed, while token accuracy measures the proportion of individual characters transcribed correctly. This distinction is important because partial matches may still be useful for record linkage, particularly when using approximate string matching algorithms.

Table 1: Transcription Performance

Field	Sequence Accuracy	Token Accuracy
First name	97.38%	98.61%
Last name	94.54%	98.29%

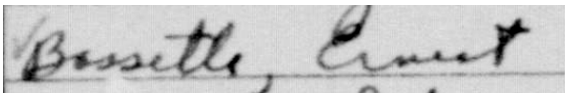
Our model achieves 97.4% sequence accuracy and 98.6% token accuracy for first names. Performance on last names is slightly lower at 94.5% sequence accuracy and 98.3% token accuracy. These accuracy rates compare favorably to previous approaches and human transcription benchmarks. Importantly, our model predicts each character independently rather than attempting to recognize whole words, which helps avoid biasing transcriptions toward common names.

However, these metrics may overstate real-world performance because our test set, like our training data, only includes cases where human transcribers initially agreed. When applying the model to cases where human transcribers disagree, performance is likely lower. This selection effect merits careful examination since these disagreement cases represent nearly 30 percent of our sample.

For records where human transcribers disagree, our model's transcription matches one of the two human transcriptions in 58% of cases, providing some validation of our approach. Interestingly, the model agrees with Ancestry's transcription in 11% of these cases and with FamilySearch's in 47% of cases, suggesting that FamilySearch transcriptions may be somewhat more reliable on average. In the remaining 42% of disagreement cases, our model produces a transcription that matches neither human transcription. A manual review of a random sample of these cases suggests that they largely fall into two categories: (a) cases involving particularly illegible handwriting, or (b) cases resulting from incorrect segmentation.

Our analysis reveals three main types of transcription errors, which we illustrate in Figure 2. Panel A shows examples of segmentation failures, where our system misidentifies the boundaries of individual cells in census tables, particularly when forms are damaged or misaligned. As shown in these examples, this can lead to incomplete name tran-

scriptions - for instance, capturing only part of a long surname that extends beyond cell boundaries. Panel B demonstrates legitimate character ambiguity, where multiple interpretations are plausible (e.g., determining whether a character is an 'a' or 'o'). Panel C demonstrates cases where our model struggles with unusual handwriting styles, showing paired examples of successful and failed transcriptions for enumerators whose writing differs substantially from our training data. These limitations suggest our method will benefit from more training data, which we are currently working on.



(a) Ambiguous characters (B"o"ssette or B"a"ssette)



(b) Same name ("Archetto, Angelo") with a successful transcription by our model (left) and an incorrect one ("Arehetto, Angelo", right)

Figure 2: Examples of transcription errors made by our model

The model's computational requirements scale efficiently with dataset size. Processing the complete Rhode Island sample of 299,880 records (approximately 0.22% of the 1940 U.S. Census) required approximately 1600 GPU-hours on standard GPU hardware. Based on these results, we estimate processing the full 1940 Census will require roughly 16500 GPU-hours. While substantial, this is well within modern computational capabilities.

3 Census Linking Data and Methods

This section describes our data sources and linking methodology. We first explain our sample construction and key variables, then detail how we incorporate our improved transcriptions into existing linking approaches.

3.1 Census Records

Our analysis uses the 1930 and 1940 U.S. Census full-count datasets provided by IPUMS (Ruggles et al., 2024). These datasets were created through a collaboration between IPUMS and two genealogical organizations, Ancestry.com and FamilySearch.org (Ruggles, 2023). We begin with 1940 to build a robust initial training dataset from records where independent transcribers agree. The rich dual-transcribed data available from 1860 to 1940 provides significant opportunities to expand our training data and further improve model performance. Our method can be applied to earlier censuses which often suffer from

worse legibility and greater transcription errors, where the potential gains from improved transcription are likely even larger due to generally less legible and standardized handwriting.

We restrict our sample to Rhode Island residents to reduce computational requirements during development of our method. Rhode Island's population in 1940 was 348,255 males, of whom 299,880 were aged 10 or above and thus potentially linkable to the 1930 census. We exclude children under 10 as they would not have appeared in the 1930 census.

Within this sample, we identify 89,218 records (29.9%) where Ancestry.com and FamilySearch.org transcriptions disagree on either first or last name or age. These records form our target set for transcription improvements. The high rate of disagreement highlights the difficulty of consistent handwriting transcription even among trained human transcribers.

The key variables used for linking include first and last names as the primary linking variables, age or birth year to establish plausible matches, and birthplace (state for those born in the US, and country for the foreign-born). Middle initials are used for validation (described below) but not for linking.

3.2 Linking Methodology

We employ the widely-used linking algorithm developed by Abramitzky et al. (2012, 2014, 2019b). This algorithm proceeds in two stages. First, it identifies potential matches based on exact name matching and birth year differences within ± 2 years. Second, it requires uniqueness within the age window to establish a definitive link.

Formally, for each 1930 record i, the algorithm:

1. Identifies all 1940 records j where:
    - First and last names match exactly
    - Birth state matches exactly
    - $|BirthYear_i - BirthYear_j| \leq 2$
2. Creates a link if and only if exactly one candidate match exists within the age window

This conservative approach prioritizes false negative over false positive errors, a common practice in the census linking literature. We maintain this standard algorithm rather than developing a new one to isolate the impact of improved transcriptions from changes in linking methodology.

Our baseline links use the Ancestry.com transcriptions distributed by IPUMS. We then replace names in our target set with the machine learning transcriptions described in Section 2 and rerun the linking algorithm. This allows us to measure the direct impact of transcription improvements on linking success.

While more elaborate linking algorithms exist (Helgertz et al., 2024; Buckles et al., 2023), we use this simpler approach for two reasons. First, it clearly demonstrates how transcription quality affects linking outcomes without conflating the effects of complex matching procedures. Second, it is widely used in applied research, making our improvements immediately relevant to current practice.

3.3 Analysis Sample

Table 2 presents summary statistics for our analysis sample, compared to males aged 10 and above across all states and those specifically in the Northeast. Rhode Island's population was more urban and had higher educational attainment than the national average, potentially affecting the generalizability of our results. However, the rate of transcription disagreement (29.9%) is similar to national levels, suggesting our findings about transcription improvements may generalize well.

The subsample where transcribers disagree shows some systematic differences from records where they agree. Disagreement is more common among foreign-born individuals and those with foreign-born parents, as well as among individuals without any formal education (see Table 3).

These patterns align with intuition about when transcription is more challenging. They also highlight the potential for transcription errors to create systematic biases in linked samples, particularly for historically disadvantaged populations.

We validate our links using middle name initials, which were recorded but not used in the linking process. This approach follows Bailey et al. (2020) and provides an independent check on link quality. The validation rate – the share of links where middle initials match when available – serves as our primary quality metric alongside raw linking rates.

A limitation of our Rhode Island sample is that we cannot observe false negatives caused by transcription errors in the 1930 census, which we have not yet processed with our machine learning pipeline. This suggests our measured improvements may understate the potential gains from applying our method to both censuses in a linking pair.

In the next section, we present results showing how our transcription improvements affect both the quantity and quality of links in this sample, with particular attention to traditionally under-linked populations.

4 Results

This section presents the impact of our improved transcriptions on census linking outcomes. We first document overall improvements in linking rates, then analyze heterogeneous effects across subgroups with particular attention to racial and geographic disparities. Finally, we validate our results using independent data sources.

Table 2: Socio-demographic characteristics of the analysis sample compared to the population and the Northeast residents

	US	Northeast	Rhode Island
Incongruent transcription	0.283	0.306	0.299
Black	0.089	0.034	0.014
American Indian	0.002	0.000	0.000
Asian	0.003	0.001	0.001
Born in Northeast	0.226	0.736	0.749
Born in Midwest	0.292	0.019	0.011
Born in South	0.318	0.036	0.010
Born in West	0.052	0.003	0.002
Foreign-born	0.111	0.203	0.227
Father is foreign-born	0.238	0.449	0.531
Live in urban area	0.555	0.728	0.714
Live on farm	0.233	0.071	0.026
No education	0.031	0.034	0.038
Graduated elementary sch.	0.631	0.698	0.654
White-collar occupation	0.264	0.320	0.281
Skilled occupation	0.330	0.407	0.494
Unskilled occupation	0.275	0.239	0.212
Farmer	0.131	0.034	0.014
Yearly income	1,424.649	1,574.490	1,515.744
Observations	55,350,482	15,358,520	298,388

This table compares the socio-demographic characteristics of three samples: all males aged 10 and above in the 1940 Census (column labeled "US"); the same demographic group residing in the Northeast ("Northeast"); and those in Rhode Island. Females and individuals under age 10 are excluded, as they are not linked by the algorithm used. All characteristics, except for yearly income, are binary.

Table 3: Characteristics of individuals with congruent transcriptions vs. incongruent transcriptions

	Transcription		Diff.
	Congruent	Incongruent	
Black	0.015	0.011	-0.004***
American Indian	0.000	0.000	0.000
Asian	0.001	0.001	0.000
Born in Northeast	0.762	0.722	-0.040***
Born in Midwest	0.009	0.009	-0.000
Born in South	0.010	0.008	-0.002***
Born in West	0.002	0.002	0.000
Foreign-born	0.216	0.258	0.042***
Father is foreign-born	0.517	0.574	0.057***
Live in urban area	0.707	0.715	0.008***
Live on farm	0.028	0.022	-0.007***
No education	0.035	0.044	0.010***
Graduated elementary sch.	0.656	0.640	-0.016***
White-collar occupation	0.277	0.288	0.011***
Skilled occupation	0.496	0.487	-0.009***
Unskilled occupation	0.212	0.213	0.001
Farmer	0.015	0.012	-0.002***
Yearly income	1,504.939	1,481.150	-23.789***
Observations	197,496	84,187	298,388

Note: The table compares the socio-demographic characteristics of individuals whose name transcriptions are congruent between Ancestry and FamilySearch with those whose transcriptions are not congruent. The sample is restricted to males in Rhode Island aged 10 and above. All characteristics, except for yearly income, are binary. The column labeled "Diff." shows the difference in means between the two samples. Standard errors are omitted. * for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$.

4.1 Impact on Linking Rates

Table 4 presents our main results comparing linking rates before and after transcription improvements. For records where human transcribers initially disagreed (our target set), the linking rate increased from 8 to 21%, representing a 147 percent improvement. This dramatic increase suggests that transcription errors were a major barrier to successful linking for these records.

Table 4: Quality of linked samples before and after transcriptions are improved

		Before		After		% change
		# of records	Rate (Share)	# of records	Rate (Share)	
Linkage rate	Target	84,538	0.084	84,538	0.208	+147%
	All	299,880	0.270	299,880	0.302	+12%
Share validated	Target	2,078	0.608	4,835	0.839	+38%
	All	21,663	0.857	24,265	0.876	+2%

Note: This table presents changes in the linkage rate and the share validated before and after transcriptions are improved. 'Before' improvement links use transcriptions from Ancestry. Share validated is equal to the number of linked records with matching middle name initials divided by the number of linked records with non-missing middle name initials.

The quality of these new links appears high. The share of links validated by matching middle initials increased from 61% to 84%, a 38 percent improvement. This simultaneous increase in both quantity and quality of links suggests that our improved transcriptions are recovering true matches rather than creating spurious ones.

When averaged across all records (including those where human transcribers agreed), our improvements increased the overall linking rate by 12 percent and the overall validation rate by 2 percent. These aggregate effects are smaller because they include the roughly 70 percent of records where human transcribers agreed and thus received no transcription improvements.

4.2 Validation Using 1% Sample

To further validate our transcription improvements, we leverage genealogical profiles from FamilySearch.org linked to the IPUMS 1% sample of the 1940 census. Following Hwang and Squires (2024), who used these profiles to study measurement error in census data, we treat the names recorded in FamilySearch profiles as reference "true" names since they incorporate information from multiple historical sources beyond just census records.

Our validation sample consists of records that satisfy three criteria: (1) they appear in the IPUMS 1% sample, (2) they have an associated FamilySearch profile with birth date information (ensuring the profile incorporates non-census sources), and (3) at least one of our three transcriptions (Ancestry, FamilySearch census, or machine learning) matches

the profile name exactly. This last restriction helps exclude cases where name differences likely reflect legitimate variations (e.g., Americanization of names) rather than transcription errors.

We expect this validation to reveal systematic differences between human and machine transcription errors. Based on our preliminary analysis of error patterns, we hypothesize that our character-by-character prediction approach produces more localized errors compared to human transcription—typically single-character substitutions or confusion between similar-looking characters like n/r or a/o. In contrast, human transcription differences tend toward more substantial variations, such as multiple-character differences or entirely missing names.

This validation approach offers key advantages because FamilySearch profile names incorporate multiple historical sources and have often been verified by descendants. While the 1% sample may not be fully representative, this validation exercise provides important evidence about whether our improvements represent genuine accuracy gains rather than arbitrary changes to difficult-to-read names. Results from this validation analysis are forthcoming and will be included in the final version of this paper.

4.3 Heterogeneous Effects

A central question is how machine learning transcription affects linking rates across different populations and document qualities. We analyze heterogeneous effects across three dimensions: demographic characteristics, geographic patterns, and document legibility. Our findings suggest that our improvements are largest where transcription challenges have historically been most severe.

Demographic Characteristics

Figure 3 shows how linking rates vary across different socio-demographic groups. The left panel displays pre-improvement linking rates, which range from 13.6% for those with no schooling to 30.9% for clerical workers, highlighting substantial variation. The right panel shows percentage increases in linking rates after our improvements.

The baseline linking rate for Black residents in our sample was 18.3% compared to 27.1% for white residents, consistent with the well-documented lower match rates for Black Americans in historical census linking (Abramitzky et al., 2019a; Bailey et al., 2020). Our improvements increased linking rates by similar rates for both groups: 11.4% for Black residents compared to 12.2% for white residents. We likewise find consistent improvements across groups with lower baseline linking rates, including foreign-born individuals (+14.5%), those with no formal schooling (+9.2%), and those with non-English mother tongue (+14.5%). The exception is institutional inmates, whose linking rate decreased slightly (-1%). This appears to reflect particular challenges in segmenting institutional census pages, which often used modified formats.

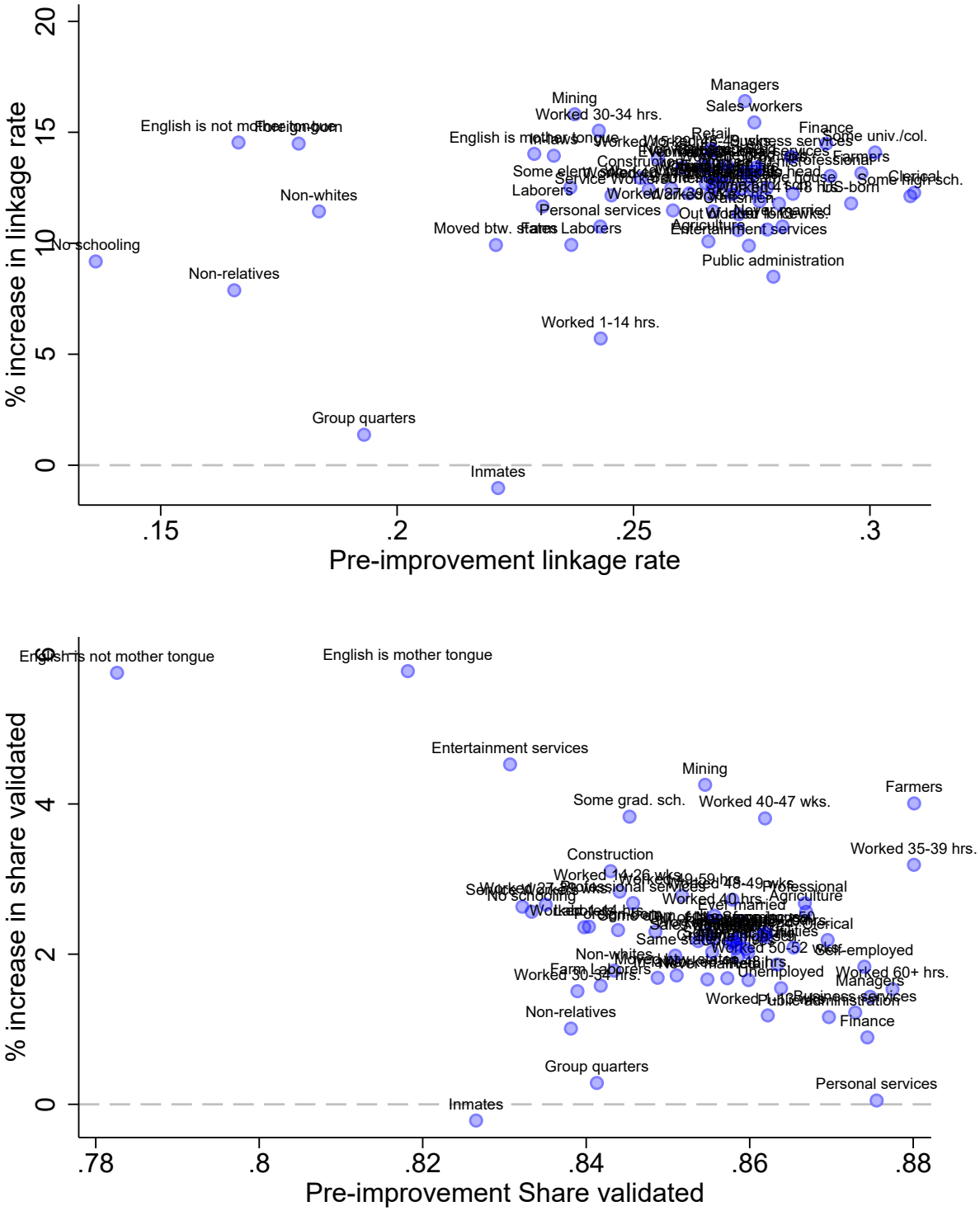


Figure 3: Quality of linked samples before and after transcription improvements for each socio-demographic group

Note: These scatter plots display the quality measures of linked samples before and after transcription improvements for each socio-demographic group. Each socio-demographic group is defined as individuals in our analysis sample who share the same socio-demographic characteristic. For example, the marker labeled "Group quarters" corresponds to people who reported in the 1940 census as living in a group quarter. We use the following socio-demographic variables to define groups: residence type, relation to head, marital status, race, birthplace, years of schooling, employment status, self-employment status, occupation, industry, weeks worked in 1939, hours worked in the previous week, 5-year migration status, and non-wage income status. The dashed line in each panel is the 45-degree line.

Geographic Patterns

Geographic variation in transcription improvements reveals important patterns in both the scope of initial transcription problems and the effectiveness of our solution. At the county level, the 5th and 95th percentile of the distribution of the share of records with incongruent transcription is 18% and 42%, with notably higher rates occurring in counties with above-median share of individuals born in the South, below age 30, or without formal education.

Document Legibility Effects

Figure 4 shows how our improvements vary with document legibility, measured at the enumeration district level following Ghosh et al. (2024). The impact is strongest for low-legibility districts, where linking rates increased by up to 35 percent. These larger gains in challenging districts reflect our model's particular effectiveness at handling hard-to-read handwriting. While human transcribers often struggle with such cases (as evidenced by their high disagreement rates), our machine learning approach can leverage patterns learned from clearer examples to decipher less legible text. This suggests that our method is most valuable precisely where traditional transcription approaches face their greatest challenges.

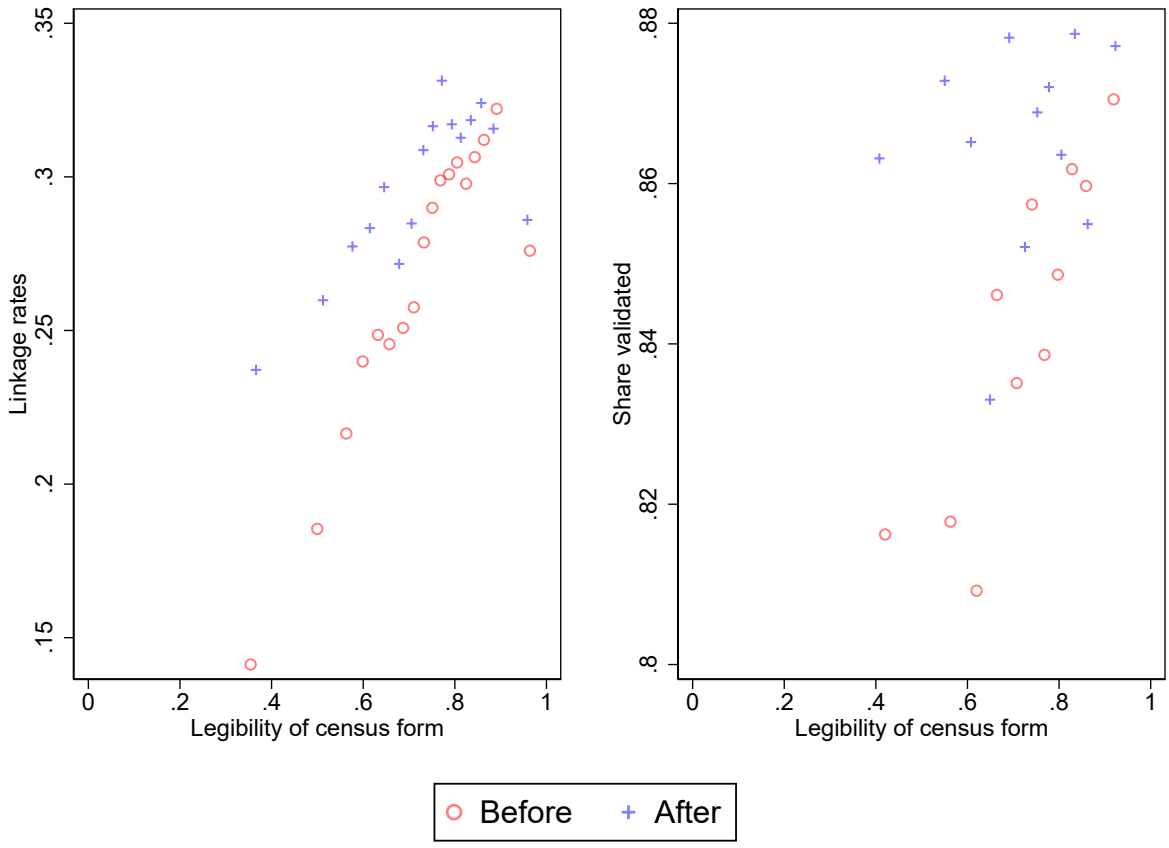
The strong performance on low-legibility records is especially valuable. Records from enumeration districts in the lowest quartile of legibility demonstrate the scope of the challenge, showing 125% higher rates of transcription disagreement and 30% lower baseline linking rates. That our approach performs best in precisely these challenging cases suggests it may help improve the representativeness of linked samples. The combination of substantial improvements for multiple demographic groups and stronger performance in low-legibility areas indicates that machine learning transcription can help expand the coverage of linked historical data.

5 Conclusion

This paper introduces a machine learning approach to improve name transcription in historical U.S. Census data, resulting in substantial gains in linking rates and match quality. Our method addresses transcription challenges in low-legibility records and provides notable improvements for historically under-linked groups such as non-English speakers, foreign-born individuals, and those with low education levels. These results not only enhance the representativeness of linked datasets but also reduce biases in the empirical findings derived from them.

Two major extensions of this work are currently in progress. First, we are scaling up our transcription pipeline to process the complete 1930 and 1940 censuses. While our Rhode Island sample demonstrates the potential of our approach, processing both censuses in their entirety will allow us to fully realize the benefits of improved transcription

Figure 4: Correlation between legibility of census form and quality of linked samples before and after transcription improvements



Note: These binscatter plots illustrate how the correlation between the legibility of census forms and quality measures of linked samples change after transcription improvements. The left panel corresponds to linkage rates, and the right panel to share validated. The unit of observation for these plots is an enumeration district. The legibility of census forms for an enumeration district is defined as in Ghosh et al. (2024), i.e., the share of records for which the transcriptions from Ancestry.com and FamilySearch.org agree. To create these plots, we follow the methodology proposed by Cattaneo et al. (2024).

for cross-census linking.

Second, we plan to directly incorporate our model's character-level uncertainty measures into the linking algorithm itself. Currently, our pipeline produces improved transcriptions that are then fed into standard linking procedures. However, our character-by-character prediction approach provides rich information about transcription uncertainty that could inform linking decisions. For example, when deciding whether two records match, the algorithm could weight character differences less heavily when they occur in positions where the model expressed low confidence in its transcription.

Beyond these immediate extensions, our approach shows promise for application to earlier U.S. censuses and international historical records, where legibility issues are often more pronounced. The method's strong performance on low-legibility records suggests it could be particularly valuable for these more challenging cases. By demonstrating that machine learning can systematically improve transcription quality while reducing demographic and geographic biases, this work represents an important step toward more representative historical microdata.

References

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson (2019b) "To the new world and back again: Return migrants in the age of mass migration," *ILR Review*, 72 (2), 300–322.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012) "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration," *American Economic Review*, 102 (5), 1832–1856.

——— (2014) "A nation of immigrants: Assimilation and economic outcomes in the age of mass migration," *Journal of Political Economy*, 122 (3), 467–506.

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James J Feigenbaum, and Santiago Pérez (2019a) "Automated linking of historical data,"Technical report, National Bureau of Economic Research.

Bailey, Martha J, Connor Cole, Morgan Henderson, and Catherine Massey (2020) "How well do automated linking methods perform? Lessons from US historical data," *Journal of Economic Literature*, 58 (4), 997–1044.

Buckles, Kasey, Adrian Haws, Joseph Price, and Haley EB Wilbert (2023) "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project,"Technical report, National Bureau of Economic Research.

Cattaneo, Matias D, Richard K Crump, Max H Farrell, and Yingjie Feng (2024) "On binscatter," *American Economic Review*, 114 (5), 1488–1514.

Cubuk, Ekin D, Barret Zoph, Jonathon Shlens, and Quoc V Le (2020) "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June.

Dell, Melissa, Jacob Carlson, Tom Bryan et al. (2024) "American stories: A large-scale structured text dataset of historical us newspapers," *Advances in Neural Information Processing Systems*, 36.

Ferrie, Joseph P (1996) "A new sample of males linked from the public use microdata sample of the 1850 US federal census of population to the 1860 US federal census manuscript schedules," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29 (4), 141–156.

Gao, Wei and Russ Tedrake (2019) "FilterReg: Robust and Efficient Probabilistic Point-Set Registration Using Gaussian Filter and Twist Parameterization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11095–11104.

Ghosh, Arkadev, Sam Il Myoung Hwang, and Munir Squires (2024) "Links and legibility: Making sense of historical US Census automated linking methods," *Journal of Business & Economic Statistics*, 42 (2), 579–590.

Goodfellow, Ian J, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet (2013) "Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks," *arXiv preprint arXiv:1312.6082*.

Helgertz, Jonas, Nesile Ozder, Steven Ruggles et al. (2024) "IPUMS Multigenerational Longitudinal Panel: Version 1.2 [dataset]," 10.18128/D016.V1.2.

Huang, Gao, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger (2016) "Deep networks with stochastic depth," in *European conference on computer vision*, 646–661, Springer.

Hwang, Sam Il Myoung and Munir Squires (2024) "Linked samples and measurement error in historical US census data," *Explorations in Economic History*, 101579.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley (2021) "Combining family history and machine learning to link historical records: The Census Tree data set," *Explorations in Economic History*, 80, 101391.

Ruggles, Steven (2023) "Collaborations Between IPUMS and Genealogical Organizations, 1999–2022," *Historical life course studies*, 13, 1.

Ruggles, Steven, Matt A. Nelson, Matthew Sobek, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Evan Roberts, and J. Robert Warren (2024) "IPUMS Ancestry Full Count Data: Version 4.0 [dataset]," 10.18128/D014.V4.0.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li (2021) "Layoutparser: A unified toolkit for deep learning based document image analysis," in *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, 131–146, Springer.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014) "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, 15 (1), 1929–1958.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna (2016) "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tan, Mingxing and Quoc Le (2021) "EfficientNetV2: Smaller Models and Faster Training," 139, 10096–10106, `https://proceedings.mlr.press/v139/tan21a.html`.

Zhong, Zhun, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang (2020) "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 13001–13008.

A Appendix: Machine Learning Transcription Methodology

A.1 Pipeline Overview

Our machine learning pipeline splits the handwritten text recognition task into three sequential components: (1) layout classification, (2) table segmentation, and (3) handwritten text recognition. For census forms, which follow a standardized layout, we can skip the first step. This appendix provides detailed technical specifications for each component of our pipeline.

The focus of our work is limited to tabular documents - specifically, those containing tables with sequences of numbers and characters organized in cells. This structure is characteristic of census forms, birth records, death records, and similar historical documents.

A.2 Table Segmentation

Table segmentation extracts images corresponding to each field (cell) in a given table from the source image. We accomplish this through the following process:

First, we note the coordinates of line intersections and line endpoints, which we refer to as template key points. We construct an 'overlay,' defined as the set of rectangles that encloses each field of interest, where each rectangle is represented by the coordinates of its four corners.

For identifying table structures, we use a semantic segmentation model to extract vertical and horizontal lines from the census records. We then align the set of intersection points of each page with pre-specified templates using efficient probabilistic point-set registration (FilterReg) by Gao and Tedrake (2019). This allows us to calculate transformation matrices for each image which we use, in combination with our pre-specified overlay, to crop each field of interest into a separate image.

The point set registration task involves aligning points between an image and a template. Our approach uses neural networks to identify lines in tables and then compares images using point cloud transformations. Once we identify the points, it becomes straightforward to crop out the segments.

Once the template and overlay are established, we process the remaining documents, which we designate as 'target' images. For each target image, we identify key points corresponding to those defined in the template, accomplished through standard computer vision operations that detect vertical and horizontal table lines.

To address the challenge of image distortion, which commonly occurs in historical documents due to scanning artifacts and paper degradation, we implement a multi-pass processing approach. Each page undergoes processing at three different scales: the full page, the top half, and the bottom half. This redundant processing strategy helps capture details that might be lost or distorted at any single scale. For each scale, we compute a distinct transformation matrix, resulting in three separate segmentations of every field on

a census record page.

The transcription process similarly leverages this multi-scale approach. We generate three independent transcriptions for each field—one from each scale—and determine the final transcription through majority voting. In cases where no majority exists, we select the transcription associated with the highest model confidence score. This ensemble approach helps mitigate errors that might arise from processing at any single scale.

Our segmentation model incorporates a quality control mechanism by estimating the alignment accuracy between the source image points and the overlay template. During the training phase, we take a conservative approach by excluding any segmented images where our model indicates uncertainty in the segmentation quality. However, we retain these uncertain segments for post-training transcription to maintain comprehensive coverage of the census records while preventing potentially noisy examples from affecting the model's training process.

A.3 Neural Network Architecture

For transcription, we train neural networks to convert the contents of each segmented image into text. Our networks are based on the EfficientNetV2 architecture (Tan and Le, 2021). We specifically use an EfficientNetV2-S variant as a balance between performance and computational requirements. This choice is motivated by EfficientNetV2's strong trade-off between accuracy and computational requirements, as demonstrated in benchmarks against many common convolutional neural networks and vision transformer architectures

Unlike many handwritten text recognition systems that use recurrent architectures, we opt for a non-recurrent architecture that is faster to train and performs well even with limited labeled data. We combine the backbone of an EfficientNetV2 model with the classification scheme proposed in Goodfellow et al. (2013), consisting of a separate classification head for each token in a sequence. In total, we use up to 40 classification heads, allowing us to transcribe any name consisting of at most 40 characters.

Our model provides probability distributions over characters for each position in the sequence, enabling us to quantify uncertainty at both the character and name level. This uncertainty quantification is particularly valuable for downstream tasks like record linkage, where we can use these confidence measures to inform linking decisions.

Our alphabet consists of the letters a-z and "space". Due to inconsistencies in source data, we do not differentiate between "space", "'", and "-", thus treating names like "Mary M", "Mary'M", and "Mary-M" as identical.

To regularize our networks and improve generalization, we employ several complementary techniques. We use dropout (Srivastava et al., 2014) with probability 0.4 to randomly disable neurons during training, preventing co-adaptation of feature detectors. Additionally, we implement stochastic depth (Huang et al., 2016) with probability 0.25, which randomly drops entire layers during training to create an implicit ensemble of networks with varying depths. Label smoothing (Szegedy et al., 2016) with value 0.1 helps

prevent the model from becoming overconfident in its predictions. We also apply weight decay at a rate of 0.000007 to control model complexity and prevent overfitting.

For data augmentation, we employ two primary techniques to artificially expand our training dataset and improve model robustness. We use RandAugment (Cubuk et al., 2020) with N=2 and M=7, which automatically searches for and applies optimal image augmentation policies. We also implement Random erase (Zhong et al., 2020) with probability 0.4, which randomly masks out rectangular regions of input images during training to improve the model's ability to handle occlusions and corrupted inputs.

Our optimization strategy combines several approaches to ensure stable and efficient training. We use stochastic gradient descent with a momentum value of 0.9 to help overcome local minima and accelerate convergence. A cosine annealing scheduler with warm-up modulates the learning rate throughout training, starting with a gradual increase and then smoothly decreasing it according to a cosine function. We apply gradient clipping at 0.02 to prevent explosive gradients, and use a batch size of 128 to balance between computational efficiency and optimization stability.

Model performance is evaluated using two metrics:

1. Sequence accuracy: measures the share of transcriptions where each character exactly matches the label
2. Token accuracy: measures the share of tokens that are transcribed correctly, allowing partial credit for names transcribed mostly correctly

Training parameters are specified as follows:

Table A.1: Model Training Parameters

Parameter	Value
RandAugment	N=2, M=7
Batch size	128
Gradient clip value	0.02
Dropout probability	0.4
Stochastic depth probability	0.25
Peak learning rate	0.5
Momentum	0.9
Random erase probability	0.4
Label smoothing	0.1
Weight decay	0.000007

A.4 Training Process

We train separate models for first and last names, motivated in part by a limitation of our training data labels, which do not distinguish between a last name being present on an

image or instead being recorded as a "ditto" mark referring to the last name in a previous row.

Our training data is constructed from human transcriptions from Ancestry and FamilySearch of the 1940 US census. To obtain high-quality labels, we only include transcriptions where both first and last names, as well as age, match exactly between the two sources. Further, we exclude all segmented images where our segmentation model reports any level of uncertainty. This results in 392,085 labels for our initial dataset, of which we reserve 10% as a test set for out-of-sample performance measurements. Our training data includes 29,641 unique first names, including differentiating between variations such as "James F" and "James W".

To mitigate issues related to "dittos" when creating the training data for our last name model, we discard all rows where the last name of the individual in the previous row matches that of the current row. A name is never written as a "ditto" on the source image unless it is the same last name as that of the above row, and we thus remove all segmented images with last names written as a "ditto". This has the side effect of also removing all "non-ditto" last names that happen to match that of the previous row, but we believe this conservative approach is nevertheless optimal, as we have observed significant challenges when training a last name transcription model where a large share of all labels do not match the content of the segmented image. After applying these restrictions, we retain 121,465 labels for last names, consisting of 28,032 unique last names, of which we again reserve 10

We evaluate model performance using two complementary metrics. "Sequence accuracy" measures the share of transcriptions where each character exactly matches the label, marking a transcription as incorrect if just one character differs from the label. "Token accuracy" measures the share of tokens that are transcribed correctly, awarding partial points for a name transcribed mostly right. This second measure may be more appropriate when using the transcribed names for linkage that allows some degree of non-exactness, such as when using Jaro-Winkler string distance.

We first train a model for transcription of first names using a pre-trained EfficientNetV2-S model as the starting point of the model's backbone and train for 90 epochs. We then use this model to initialize our model for transcription of last names, which we train for 180 epochs (we train for additional epochs due to the smaller number of training samples). Our first name model achieves 97.38% sequence accuracy and 98.61% token accuracy, while our last name model achieves 94.54% sequence accuracy and 98.29% token accuracy. The lower sequence accuracy for last names may reflect the smaller number of training samples compared to our first name model.

While these transcription accuracies are very high, this may in part reflect that the images we use to evaluate our models are relatively easy to transcribe, as we only include labels where Ancestry and FamilySearch transcriptions agree and further exclude images that may be incorrectly segmented. This likely upwards biases the transcription accuracies we report, but avoiding this bias without introducing a potential downwards bias is challenging. While keeping only labels where both transcriptions agree may lead to an

overrepresentation of “easy” images, removing this restriction would require choosing either the Ancestry or FamilySearch label as ground truth, potentially including a large share of incorrect labels.