Enhancing the Precision of Vast Historical Datasets via Deep Learning and AI: A Case Study on Mitigating Misinterpretation Challenges within the 1950 U.S. Census

Mengyue Zhao

MSc. Economic and Social History, University of Oxford, Class of 2019

beczhaozmy0@gmail.com

Abstract

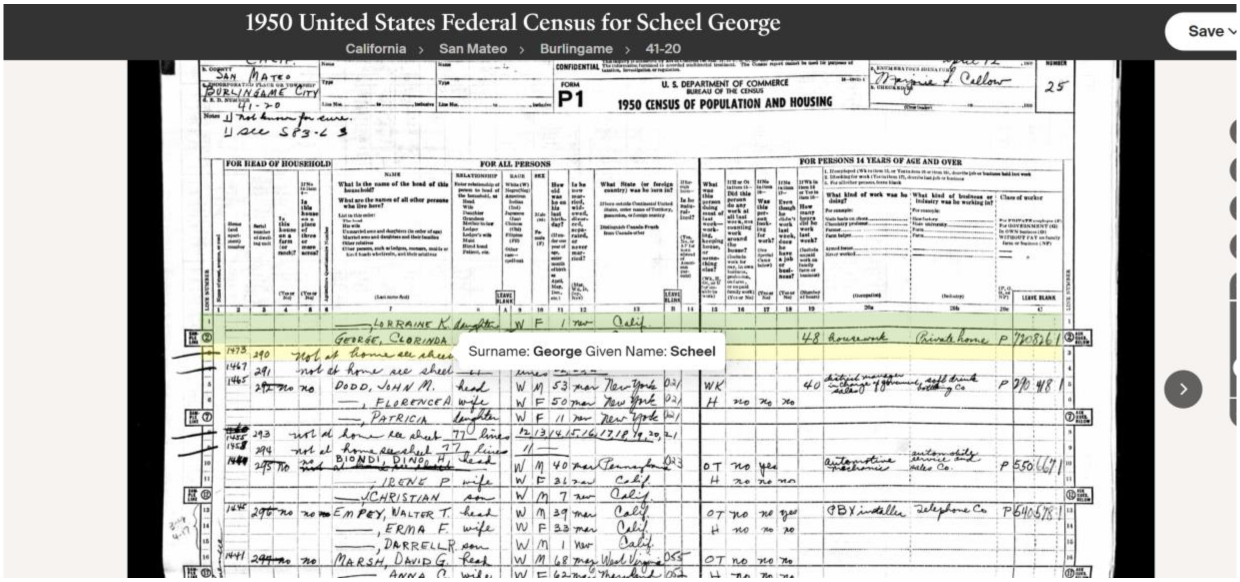
The integrity and reliability of historical records have emerged as critical concerns for researchers in the humanities and social sciences. Traditional methodologies often fall short in addressing the complexities and nuances inherent in historical documents, leading to transcription errors and misinterpretations that can significantly skew data analysis. This study introduces a dual-model deep learning framework that combines semantic segmentation and transfer learning to accurately identify and correct transcription errors across 15 million pages of U.S. Census records. Validated in collaboration with the Minnesota Population Center, the experiment achieved a 95% accuracy rate on a test set of 56,310 entries, effectively eliminating up to 1-2 million erroneous entries and enhancing the trustworthiness of historical data for research applications.

Introduction

This study applies cutting-edge deep learning and AI models to improve data accuracy across 15 million pages of U.S. Census records by identifying and correcting transcription errors and erroneous entries. Through a human-validated test set of 56,310 entries, the project achieved 95% accuracy in correcting transcription errors, eliminating an estimated 1-2 million errors. This sets a new benchmark for integrating technology into social sciences, with broad implications for improving research reliability in any endeavor requiring large-scale data collection from historical sources.

Problem Statement

The 1950 census shows an estimated 1-2 million overcounted entries across 15 million pages when cross-referenced with county-level data. Traditional OCR methods struggle with these inaccuracies, such as confusing "Noah" with "No One." My project, conducted in partnership with the Minnesota Population Center, utilizes deep learning to address these inaccuracies, achieving a 95% accuracy rate in error detection and correction, ultimately enhancing the reliability of historical data for research. In the example bellow, row 3 says “not at home see sheet [...]” but Ancestry.com detects “George Scheel” based on visual similarity.



1950 United States Federal Census for Scheel George

California › San Mateo › Burlingame › 41-20

SAN MATEO
BURLINGAME CITY
41-20
Notes 1] Not known for cure.
1] see S83-L3

CONFIDENTIAL
FORM P1
U. S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
1950 CENSUS OF POPULATION AND HOUSING
Marjorie F. Callow
25

FOR HEAD OF HOUSEHOLD | FOR ALL PERSONS | FOR PERSONS 14 YEARS OF AGE AND OVER

Surname: **George** Given Name: **Scheel**

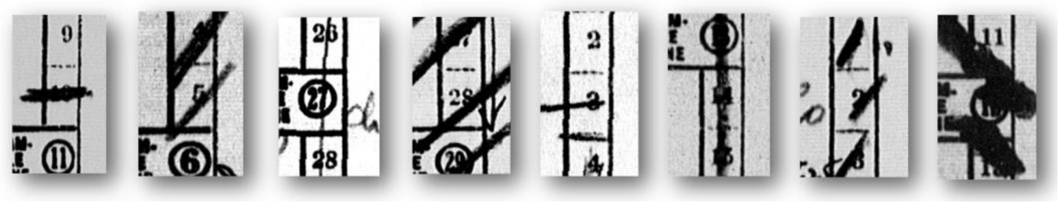
Line	2	3	4	5	7	8	9	10	11	12	13	B	15	16	17	18	19	20a	20b	20c	
1					—, LORRAINE K.	daughter	W	F	1	nev	Calif.										
2					GEORGE, CLORINDA												48	housework	Private home	P	720826 1
3	1473	290			not at home see sheet																
4	1467	291			not at home see sheet																
5	1465	~~292~~	no	no	DODD, JOHN M.	head	W	M	53	mar	New York	021	WK				40	district manager in charge of government sales	soft drink bottling Co	P	290 418 1
6					—, FLORENCE A.	wife	W	F	50	mar	New York	021	H	no	no	no					
7					—, PATRICIA	daughter	W	F	11	nev	New York	021									
8	~~1460~~ 1455	293			not at home see sheet 77 lines		12, 13, 14, 15, 16, 17, 18, 19, 20, 21														
9	1458	294			not at home see sheet 77 lines		11 —														
10	~~1444~~	295	no	~~no~~	~~not at home see sheet~~ BIONDI, DINO H.	head	W	M	40	mar	Pennsylvania	023	OT	no	yes			automotive mechanic	automobile service and sales Co.	P	550 667 1
11					—, IRENE P	wife	W	F	31	mar	Calif.		H	no	no	no					
12					—, CHRISTIAN	son	W	M	7	nev	Calif.										
13	1445	296	no	no	EMPEY, WALTER T.	head	W	M	39	mar	Calif.		OT	no	no	yes		PBX installer	Telephone Co	P	640 578 1
14					—, ERMA F.	wife	W	F	33	mar	Calif.		H	no	no	no					
15					—, DARRELL R.	son	W	M	1	nev	Calif.										
16	1441	~~297~~	no	no	MARSH, DAVID G.	head	W	M	68	mar	West Virginia	055	OT	no	no	no					
17					—, ANNA C.	wife	W	F	62	mar	Maryland	052	H	no	no	no					

Limitations of Traditional Computer Vision for Complex Historical Documents

Edge detection and traditional computer vision methods are poorly suited for processing the kind of complex information seen in historical census documents like those in the image. These records contain dense, overlapping elements, such as strikethroughs, circles, and crossed-out numbers, which would confuse an edge detection algorithm. Such algorithms rely on detecting sharp contrasts and well-defined lines, but the handwritten marks and varied structures in these documents do not fit such a pattern. Overlapping strokes and faded lines are common, further complicating the task. The inconsistent handwriting, varied thickness of lines, and degradation artifacts introduce noise that would likely be misinterpreted as valid edges or boundaries, producing inaccurate results. Without a semantic understanding of the content, edge detection would fail to distinguish between meaningful data points and irrelevant markings, leading to errors and missed information.

Moreover, the structure of historical census records lacks the standardization that edge detection requires. Each entry may contain unique notations, symbols, and inconsistent row structures, making it challenging for basic algorithms to generalize. Edge detection alone would struggle to parse these subtle distinctions, as it lacks the ability to classify elements within the context of a historical document's layout. In contrast, a deep learning framework equipped with semantic segmentation and image classification can be trained to recognize and interpret these complex patterns and contextual nuances, enabling it to differentiate between data points, noise, and layout elements. This approach allows for a higher accuracy rate, as it can adapt to the specific irregularities in each record, handling the variability in handwriting, formatting, and document quality that edge detection cannot manage effectively.

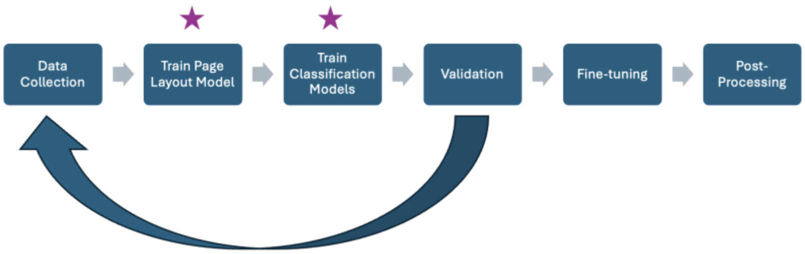
A subset of examples accurately identified by the Deep Learning Framework:



Dual Deep Learning AI Framework for Precision Data Correction

The Dual Deep Learning AI Framework, honored with the Digital Innovation Award at the 2022 World Economic History Congress, integrates two specialized models: a page segmentation model and a classification vision model/OCR engine. This dual-model framework extracts and corrects data from large, unstructured historical datasets like census records with high precision, offering unprecedented reliability for research dependent on vast primary sources.

Dual Deep-learning AI framework Iteration Process



This diagram outlines the iterative process of developing a machine learning framework for historical document analysis, showing the key stages involved from data collection to post-processing. Here’s a breakdown of each step:

1. Data Collection: Gather a large set of census images with accurate labels, prioritizing pages with strikethroughs for comprehensive training.

2. Train Page Layout Model: Train a model to segment uniform census structures, focusing on rows, columns, and marked areas for downstream classification.

3. Train Classification Models: Develop models to identify specific content and patterns, particularly strikethroughs, within segmented areas.

4. Validation: Validate using a test set of 1,877 pages with human-verified data, with a bias in favor of pages with a lot of strike-throughs.

5. Fine-Tuning: Refine models by focusing on challenging or special cases.

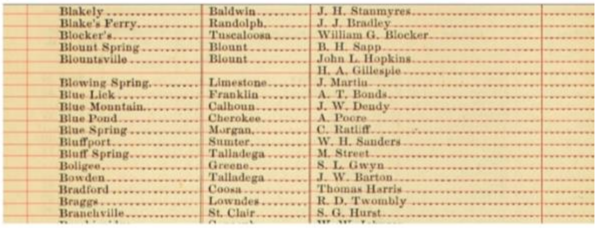
6. Post-Processing: Build a pipeline to format outputs into human-readable data, ensuring clean, usable results.

7. Iteration (Feedback Loop): Continuously improve by feeding insights from post-processing back into data collection and training.

Intelligent Page Segmentation Model

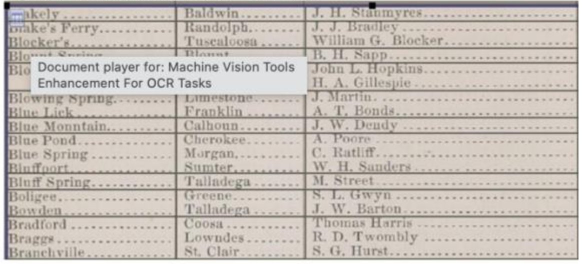
Page Segmentation is a critical yet often overlooked step when working with historical documents. While commercial or open-source OCR software performs well on modern, clearly structured tables, it struggles with the unique complexities of historical formats. The example below demonstrates the advantage of a customized solution designed for historical data, like the *Official Register of the United States*. Here, OmniPage fails to accurately detect rows due to the dotted line structure, which renders the resulting data unreliable for downstream use.

Figure 2. ROI Detection Results -- Customized OCR Tool



Blakely	Baldwin	J. H. Stanmyres
Blake's Ferry	Randolph	J. J. Bradley
Blocker's	Tuscaloosa	William G. Blocker
Blount Spring	Blount	B. H. Sapp
Blountsville	Blount	John L. Hopkins
		H. A. Gillespie
Blowing Spring	Limestone	J. Martin
Blue Lick	Franklin	A. T. Bonds
Blue Mountain	Calhoun	J. W. Dendy
Blue Pond	Cherokee	A. Poore
Blue Spring	Morgan	C. Ratliff
Bluffport	Sumter	W. H. Sanders
Bluff Spring	Talladega	M. Street
Boligee	Greene	S. L. Gwyn
Bowden	Talladega	J. W. Barton
Bradford	Coosa	Thomas Harris
Braggs	Lowndes	R. D. Twombly
Branchville	St. Clair	S. G. Hurst

Figure 3. ROI Detection Results -- OmniPage



Document player for: Machine Vision Tools Enhancement For OCR Tasks

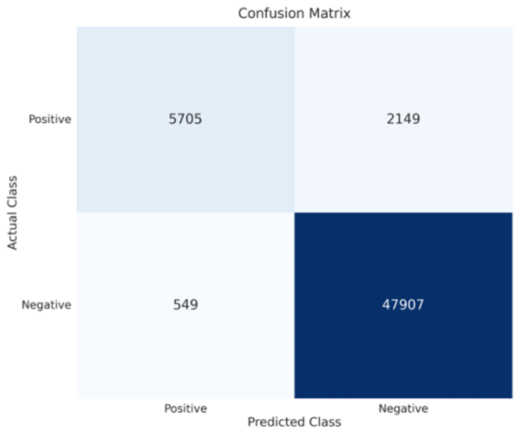
akely	Baldwin	J. H. Stanmyres
ake's Ferry	Randolph	J. J. Bradley
Blocker's	Tuscaloosa	William G. Blocker
Blo		B. H. Sapp
Blo		John L. Hopkins H. A. Gillespie
Blowing Spring Blue Lick	Limestone Franklin	J. Martin A. T. Bonds
Blue Mountain	Calhoun	J. W. Dendy
Blue Pond Blue Spring Bluffport	Cherokee Morgan Sumter	A. Poore C. Ratliff W. H. Sanders
Bluff Spring	Talladega	M. Street
Boligee Bowden	Greene Talladega	S. L. Gwyn J. W. Barton
Bradford Braggs Branchville	Coosa Lowndes St. Clair	Thomas Harris R. D. Twombly S. G. Hurst

Ensuring Trustworthy Results with Precision–Recall Balance

Balancing error detection and valid entry retention is essential in data validation. In this census project, preserving valid data was prioritized, even if a few erroneous entries were retained. This approach maintains high data integrity, as demonstrated by the model's confusion matrix. Despite the dataset's imbalance, the model reliably identifies both valid and erroneous entries, supporting the project's goal of maintaining a high standard of data trustworthiness.

The model strikes a careful balance between retaining valid data (negative class) and detecting errors (positive class). With a high specificity of 95.71%, it effectively keeps important records intact, supporting the project's data integrity goals. Additionally, the model's recall rate of 91.22% ensures it captures most errors accurately, while a precision of 72.63% indicates some tolerance for false positives, which can be easily removed by downstream processes. Overall, the model maintains a reliable dataset by preserving most valid entries while effectively identifying errors.

Confusion Matrix



Generative AI Model vs. Deep Learning Framework: Performance Comparison

Experiment Design:

This experiment tested the Gemini-1.5-pro model's ability to detect crossed-out line numbers in 1950 U.S. Census forms. Gemini-1.5-pro is one of the most performant models for multi-modal tasks on the market, it provides good baseline understanding of Large Language Models' performance on such tasks. Using a few-shot example approach, the model was tasked with identifying visibly crossed-out rows across 10 pages, where rows are marked by a line, stroke, or scribble through the line number.

Prompt Details:

“Context: The image shows a section of a 1950 U.S. Census form, with line numbers listed in the far-left column. Rows may be crossed-out to indicate corrections, removals, or other changes in the census data.

Task: Identify and return a list of line numbers that are visibly crossed-out in the provided image of the 1950 U.S. Census form. A row is considered crossed-out if there is a mark, such as a line, stroke, or scribble made with a pen, pencil, or marker, striking through or across the entire line number in the leftmost column.”

Example Image:

See Appendix A.

Example LLM Output:

10, 11

Analysis of Output:

The actual strikethrough spans lines 8 to 30, meaning all rows within this range should have been marked as crossed-out. By only identifying lines 10 and 11, the model missed a substantial portion of crossed-out lines, indicating LLM’s limitations in detecting extended patterns and composite layouts in historical records.

Summary:

The experiment, which included a few-shot example approach and testing across 10 pages, suggests that while Gemini-1.5-pro might perform better with additional few-shot examples or fine-tuning, its current output does not look promising for complex pattern recognition in historical document layouts. The model’s limitations in handling composite page-layout and pattern analysis tasks are evident. In contrast, my deep learning framework, which integrates semantic segmentation and contextual analysis, is better suited to accurately detecting intricate patterns and preserving high data integrity within archival records.

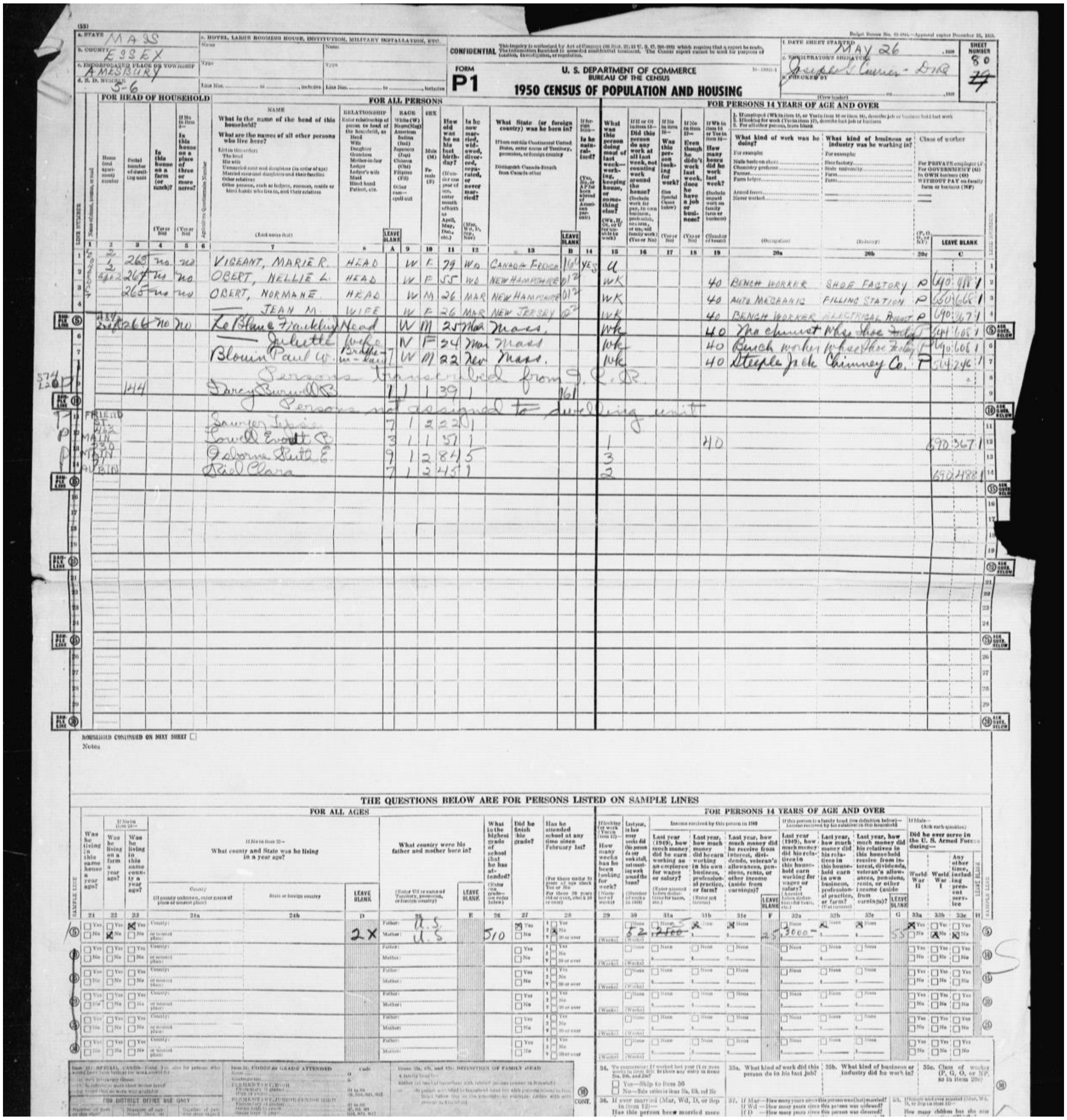
Conclusion

This research highlights the transformative potential deep learning in advancing social sciences and humanities research. By leveraging advanced AI for historical document analysis, this study significantly improves the reliability of historical records, contributing to broader discussions on identity, kinship, and community in the digital age. The project establishes new benchmarks for integrating AI in historical research, deepening our understanding of historical records' relevance to modern analysis. While multi-modal LLMs show promise, they are not yet a replacement for conventional deep-learning vision techniques, as they struggle with composite page layouts and complex pattern analysis.

Bibliography

Ancestry.com. *1950 United States Federal Census*. Lehi, UT: Ancestry.com Operations, Inc., 2022.

Appendix A



(55)

a. STATE: MASS
b. COUNTY: ESSEX
c. INCORPORATED PLACE OR TOWNSHIP: AMESBURY
d. E. D. NUMBER: 5-6

e. HOTEL, LARGE ROOMING HOUSE, INSTITUTION, MILITARY INSTALLATION, ETC.

CONFIDENTIAL

FORM P1

U. S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

1950 CENSUS OF POPULATION AND HOUSING

1. DATE SHEET STARTED: MAY 26, 1950
2. ENUMERATOR'S SIGNATURE: Joseph G. Currier - DRQ
3. CHECKED BY

SHEET NUMBER: 80 ~~79~~

FOR HEAD OF HOUSEHOLD — FOR ALL PERSONS — FOR PERSONS 14 YEARS OF AGE AND OVER

Line	2	3	4	5	6	7	8	A	9	10	11	12	13	B	14	15	16	17	18	19	20a	20b	20c	C
1	2	265	no	no		VIGEANT, MARIE R.	HEAD		W	F	79	WD	CANADA FRENCH	16	YES	U								
2	1/2 apt 2	264	no	no		OBERT, NELLIE L.	HEAD		W	F	55	WD	NEW HAMPSHIRE	012		WK				40	BENCH WORKER	SHOE FACTORY	P	690 488 1
3		265	no	no		OBERT, NORMAN E.	HEAD		W	M	26	MAR	NEW HAMPSHIRE	012		WK				40	AUTO MECHANIC	FILLING STATION	P	650 668 1
4						JEAN M.	WIFE		W	F	26	MAR	NEW JERSEY	02		WK				40	BENCH WORKER	ELECTRICAL PLANT	P	690 367 1
5	48 1/2 2nd fl	266	no	no		LeBlanc Franklin	Head		W	M	25	Mar	Mass.			Wk				40	Machinist	Whse. Shoe Fcty	P	344 605 1
6						Juliette	Wife		W	F	24	Mar	Mass			Wk				40	Bench worker	Whse Shoe Fcty	P	690 606 1
7						Blouin Paul W.	Brother-in-law	7	W	M	22	Nev	Mass.			Wk				40	Steeple Jack	Chimney Co.	P	514 246 1
8						Persons transcribed from I. C. R.																		
9		144				Darcy Burwell B.		7	1	1	39	1		161										
10						Persons not assigned to dwelling unit																		
11	FRIEND ST.					Sawyer Lizzie		7	1	2	22	1												
12	442 MAIN					Lowell Everett B.		3	1	1	57	1				1				40				690 367 1
13	230 MAIN					Osborne Ruth E.		9	1	2	84	5				3								
14	97 AUBIN					Riel Clara		7	1	2	45	1				2								690 488 1

574
126

HOUSEHOLD CONTINUED ON NEXT SHEET □

Notes

THE QUESTIONS BELOW ARE FOR PERSONS LISTED ON SAMPLE LINES

FOR ALL AGES — FOR PERSONS 14 YEARS OF AGE AND OVER

Sample line	21	22	23	24a	24b	D	25	E	26	27	28	29	30	31a	31b	31c	F	32a	32b	32c	G	33a	33b	33c
5	□ Yes □ No	□ Yes ☒ No	☒ Yes □ No			2X	Father: U.S. Mother: U.S.		510	☒ Yes □ No	☒ No		52	2500	☒ None	☒ None	25	3000-	☒ None	☒ None	55	☒ Yes	☒ No	☒ No
10	□ Yes □ No	□ Yes □ No	□ Yes □ No							□ Yes □ No				□ None	□ None	□ None		□ None	□ None	□ None		□ Yes □ No	□ Yes □ No	□ Yes □ No
15	□ Yes □ No	□ Yes □ No	□ Yes □ No							□ Yes □ No				□ None	□ None	□ None		□ None	□ None	□ None		□ Yes □ No	□ Yes □ No	□ Yes □ No
20	□ Yes □ No	□ Yes □ No	□ Yes □ No							□ Yes □ No				□ None	□ None	□ None		□ None	□ None	□ None		□ Yes □ No	□ Yes □ No	□ Yes □ No
25	□ Yes □ No	□ Yes □ No	□ Yes □ No							□ Yes □ No				□ None	□ None	□ None		□ None	□ None	□ None		□ Yes □ No	□ Yes □ No	□ Yes □ No
30	□ Yes □ No	□ Yes □ No	□ Yes □ No							□ Yes □ No				□ None	□ None	□ None		□ None	□ None	□ None		□ Yes □ No	□ Yes □ No	□ Yes □ No