Black Freedom Mapping: Application of Artificial Intelligence (AI) to Construct a National Database of Black Mobility During the Jim Crow Era

UCSF Weill Institute for Neurosciences
Memory and Aging Center

Sydney Moseley, BS[1], Pedro Pinheiro-Chagas[1,2,3], PhD, Muriel Taks Calle, MA[4], Paris B. Adkins-Jackson, PhD, MPH[4,5], & Tanisha G. Hill-Jarrett, PhD. [1,6]

[1]Memory and Aging Center, Department of Neurology, University of California San Francisco, [2]Bakar Computational Health Sciences Institute, University of California, San Francisco, [3]Center for Intelligent Imaging (ci2), University of California, San Francisco, [4]Department of Epidemiology, Mailman School of Public Health, Columbia University, [5]Department of Sociomedical Sciences, Mailman School of Public Health, Columbia University, [6]Global Brain Health Institute, University of California San Francisco

Background

Historical Context:

- During the Jim Crow era, traveling while Black meant possible life-threatening danger due to widespread racial discrimination and the threat of violence in sundown towns.
- The Motorist's Green Book, published annually from 1937 to 1964 by Harlem postal worker Victor Green, served as a vital guide for Black travelers seeking safe spaces—i.e., businesses and residences—while on the road.

Dataset Importance:

- While research on how structural racism shapes the health of diverse populations has grown, there remains critical data gaps that prohibit the use of historical data to examine associations between structural racism and population health.

Objective:

- By extracting addresses from digitized editions of The Green Book, we aim to provide a novel database of Black mobility, while describing a reproducible method for context-aware data extraction.

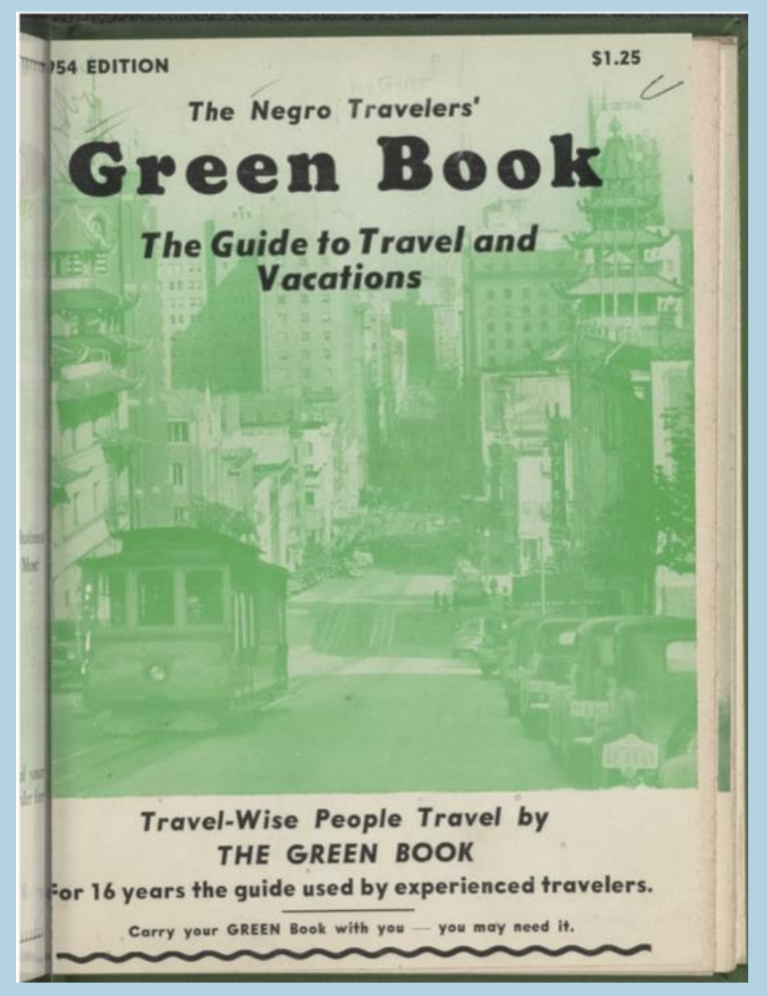
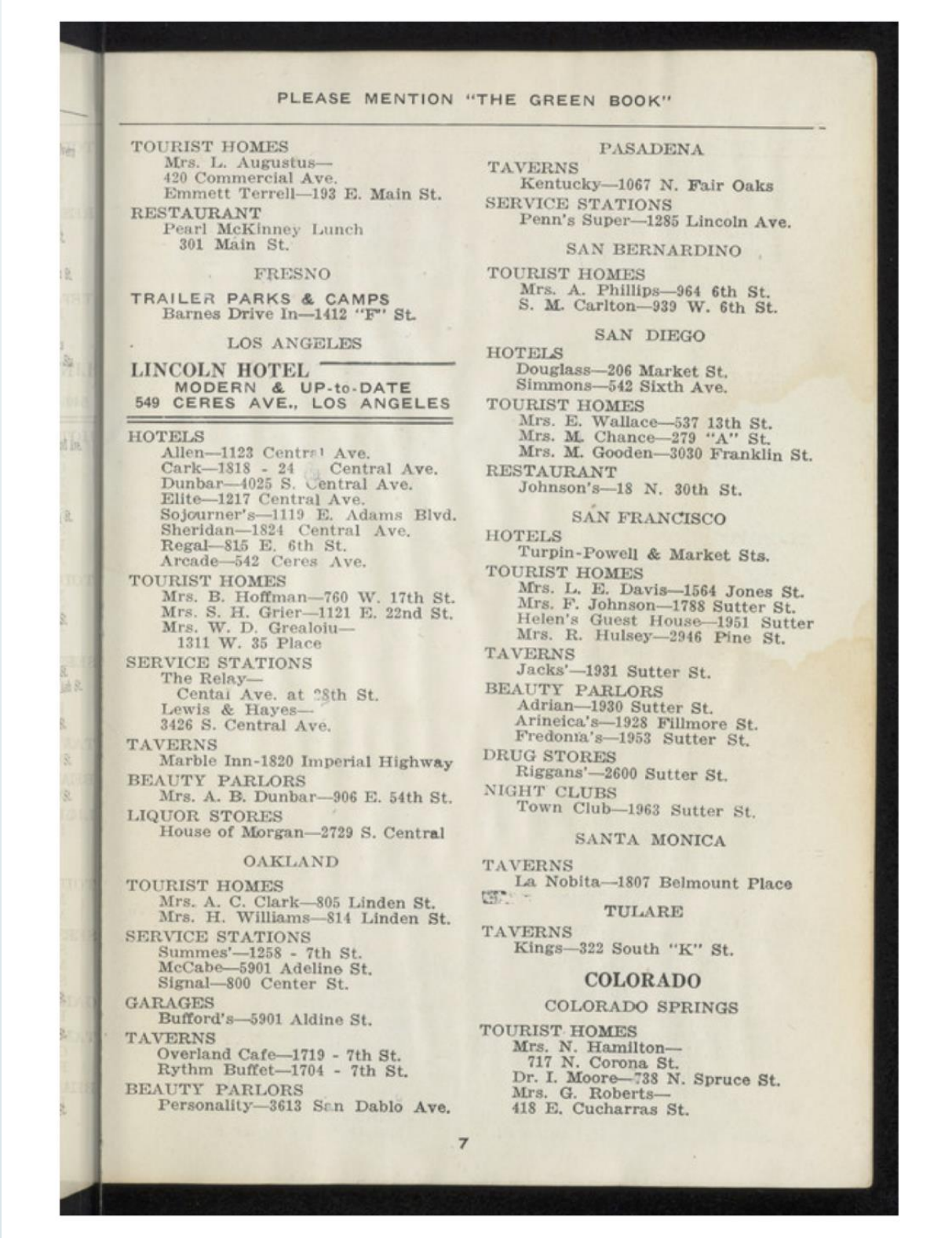

54 EDITION
$1.25
The Negro Travelers'
Green Book
The Guide to Travel and Vacations
Travel-Wise People Travel by
THE GREEN BOOK
For 16 years the guide used by experienced travelers.
Carry your GREEN Book with you — you may need it.

Figure 1. 1954 Edition Cover



PLEASE MENTION "THE GREEN BOOK"

TOURIST HOMES
Mrs. L. Augustus—
420 Commercial Ave.
Emmett Terrell—193 E. Main St.
RESTAURANT
Pearl McKinney Lunch
301 Main St.

FRESNO

TRAILER PARKS & CAMPS
Barnes Drive In—1412 "F" St.

LOS ANGELES

LINCOLN HOTEL
MODERN & UP-to-DATE
549 CERES AVE., LOS ANGELES

HOTELS
Allen—1123 Centrl Ave.
Cark—1818 - 24 Central Ave.
Dunbar—4025 S. Central Ave.
Elite—1217 Central Ave.
Sojourner's—1119 E. Adams Blvd.
Sheridan—1824 Central Ave.
Regal—815 E. 6th St.
Arcade—542 Ceres Ave.
TOURIST HOMES
Mrs. B. Hoffman—760 W. 17th St.
Mrs. S. H. Grier—1121 E. 22nd St.
Mrs. W. D. Grealoiu—
1311 W. 35 Place
SERVICE STATIONS
The Relay—
Centai Ave. at 28th St.
Lewis & Hayes—
3426 S. Central Ave.
TAVERNS
Marble Inn-1820 Imperial Highway
BEAUTY PARLORS
Mrs. A. B. Dunbar—906 E. 54th St.
LIQUOR STORES
House of Morgan—2729 S. Central

OAKLAND

TOURIST HOMES
Mrs. A. C. Clark—805 Linden St.
Mrs. H. Williams—814 Linden St.
SERVICE STATIONS
Summes'—1258 - 7th St.
McCabe—5901 Adeline St.
Signal—800 Center St.
GARAGES
Bufford's—5901 Aldine St.
TAVERNS
Overland Cafe—1719 - 7th St.
Rythm Buffet—1704 - 7th St.
BEAUTY PARLORS
Personality—3613 San Dablo Ave.

PASADENA

TAVERNS
Kentucky—1067 N. Fair Oaks
SERVICE STATIONS
Penn's Super—1285 Lincoln Ave.

SAN BERNARDINO

TOURIST HOMES
Mrs. A. Phillips—964 6th St.
S. M. Carlton—939 W. 6th St.

SAN DIEGO

HOTELS
Douglass—206 Market St.
Simmons—542 Sixth Ave.
TOURIST HOMES
Mrs. E. Wallace—537 13th St.
Mrs. M. Chance—279 "A" St.
Mrs. M. Gooden—3030 Franklin St.
RESTAURANT
Johnson's—18 N. 30th St.

SAN FRANCISCO

HOTELS
Turpin-Powell & Market Sts.
TOURIST HOMES
Mrs. L. E. Davis—1564 Jones St.
Mrs. F. Johnson—1788 Sutter St.
Helen's Guest House—1951 Sutter
Mrs. R. Hulsey—2946 Pine St.
TAVERNS
Jacks'—1931 Sutter St.
BEAUTY PARLORS
Adrian—1930 Sutter St.
Arineica's—1928 Fillmore St.
Fredonia's—1953 Sutter St.
DRUG STORES
Riggans'—2600 Sutter St.
NIGHT CLUBS
Town Club—1963 Sutter St.

SANTA MONICA

TAVERNS
La Nobita—1807 Belmount Place

TULARE

TAVERNS
Kings—322 South "K" St.

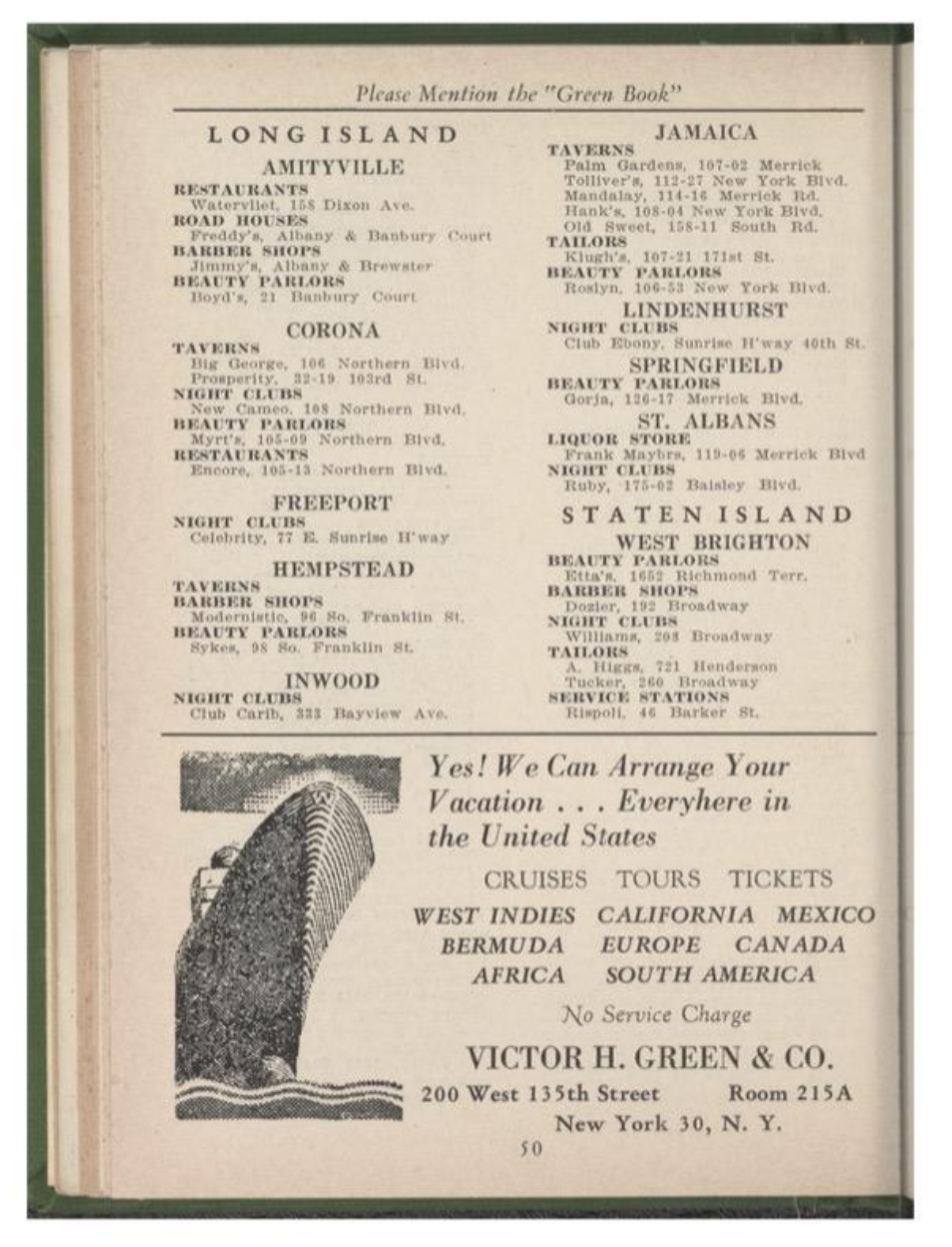
COLORADO

COLORADO SPRINGS

TOURIST HOMES
Mrs. N. Hamilton—
717 N. Corona St.
Dr. I. Moore—738 N. Spruce St.
Mrs. G. Roberts—
418 E. Cucharras St.

7

Figure 2. Page from the 1940 Green Book



Please Mention the "Green Book"

LONG ISLAND

AMITYVILLE

RESTAURANTS
Watervliet, 158 Dixon Ave.
ROAD HOUSES
Freddy's, Albany & Banbury Court
BARBER SHOPS
Jimmy's, Albany & Brewster
BEAUTY PARLORS
Boyd's, 21 Banbury Court

CORONA

TAVERNS
Big George, 106 Northern Blvd.
Prosperity, 32-19 103rd St.
NIGHT CLUBS
New Cameo, 108 Northern Blvd.
BEAUTY PARLORS
Myrt's, 105-09 Northern Blvd.
RESTAURANTS
Encore, 105-13 Northern Blvd.

FREEPORT

NIGHT CLUBS
Celebrity, 77 E. Sunrise H'way

HEMPSTEAD

TAVERNS
BARBER SHOPS
Modernistic, 96 So. Franklin St.
BEAUTY PARLORS
Sykes, 98 So. Franklin St.

INWOOD

NIGHT CLUBS
Club Carib, 333 Bayview Ave.

JAMAICA

TAVERNS
Palm Gardens, 107-02 Merrick
Tolliver's, 112-27 New York Blvd.
Mandalay, 114-16 Merrick Rd.
Hank's, 108-04 New York Blvd.
Old Sweet, 158-11 South Rd.
TAILORS
Klugh's, 107-21 171st St.
BEAUTY PARLORS
Roslyn, 106-53 New York Blvd.

LINDENHURST

NIGHT CLUBS
Club Ebony, Sunrise H'way 40th St.

SPRINGFIELD

BEAUTY PARLORS
Gorja, 126-17 Merrick Blvd.

ST. ALBANS

LIQUOR STORE
Frank Maybrs, 119-06 Merrick Blvd
NIGHT CLUBS
Ruby, 175-02 Baisley Blvd.

STATEN ISLAND

WEST BRIGHTON

BEAUTY PARLORS
Etta's, 1652 Richmond Terr.
BARBER SHOPS
Dozier, 192 Broadway
NIGHT CLUBS
Williams, 208 Broadway
TAILORS
A. Higgs, 721 Henderson
Tucker, 260 Broadway
SERVICE STATIONS
Rispoli, 46 Barker St.

Yes! We Can Arrange Your Vacation . . . Everyhere in the United States

CRUISES TOURS TICKETS

WEST INDIES CALIFORNIA MEXICO
BERMUDA EUROPE CANADA
AFRICA SOUTH AMERICA

No Service Charge

VICTOR H. GREEN & CO.
200 West 135th Street Room 215A
New York 30, N. Y.

50

Figure 3. Page from the 1954 Green Book

Method

Introduction:

- In order to create this database, one would typically manually convert the Green Books to text, an incredibly time-consuming task, or implement Optical Character Recognition (OCR).
- OCR uses pattern matching or feature extraction algorithms to convert images of text characters into a machine-readable format.
- While traditional OCR is imprecise in extracting unstructured and nonregular expressions from images of varying quality, this study introduces a novel application of Artificial Intelligence (AI) to extract historical addresses in inconsistent formats.
- We use a multimodal large language model (MLLM), which is an AI model that extends the capabilities of LLMs by processing multiple types of data, such as images, text, and audio.

Method [cont'd]

Utilizing AI:

- To extract the addresses from the Green Books, we implemented structured prompting and knowledge retrieval prompting to extract and standardize historical address data from scanned PDFs.
- PDFs of the books were converted to images, then passed through GPT-4o, a multimodal large language model (MLLM), used to extract addresses from twenty-one Green Books and convert them to text.
- A second call to the model used structured outputs to constrain the responses to our specified JSON schema.

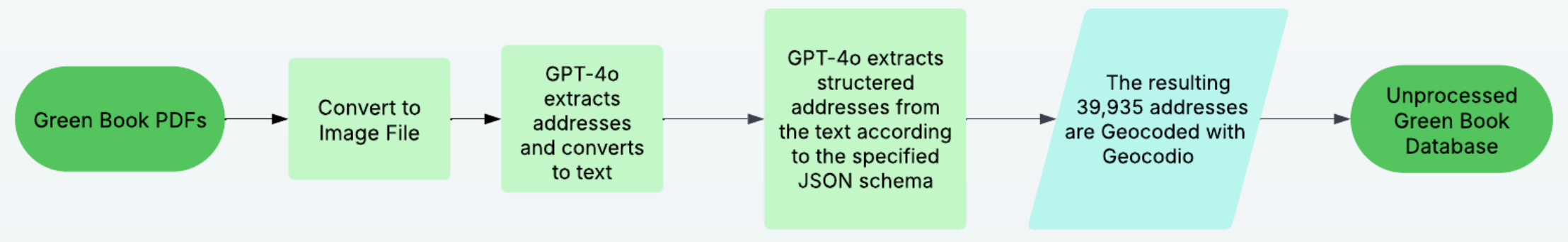

Green Book PDFs
Convert to Image File
GPT-4o extracts addresses and converts to text
GPT-4o extracts structered addresses from the text according to the specified JSON schema
The resulting 39,935 addresses are Geocoded with Geocodio
Unprocessed Green Book Database

Figure 4. Generalized workflow of using GPT-4.o to extract addresses

Geocoding and Post-Processing:

- Forward geocoding of the extracted Green Book addresses was completed in Geocodio, creating latitude and longitudinal coordinates for each address. These geospatial coordinates were linked to contemporary census data based on the 2000 U.S. Census.
- Accuracy Check: Fuzzy string matching with a threshold of 60% similarity was used as the accuracy cutoff for comparison between Geocodio's parcellation and the extracted address.
- With the remaining addresses not identified in Geocodio, Python pandas and geopy libraries were used to identify coordinates. These coordinates were matched to the U.S. Census Bureau's 2024 TIGER data using Nominatim.
- The remaining missing coordinates were manually examined for error types and corrected if applicable.

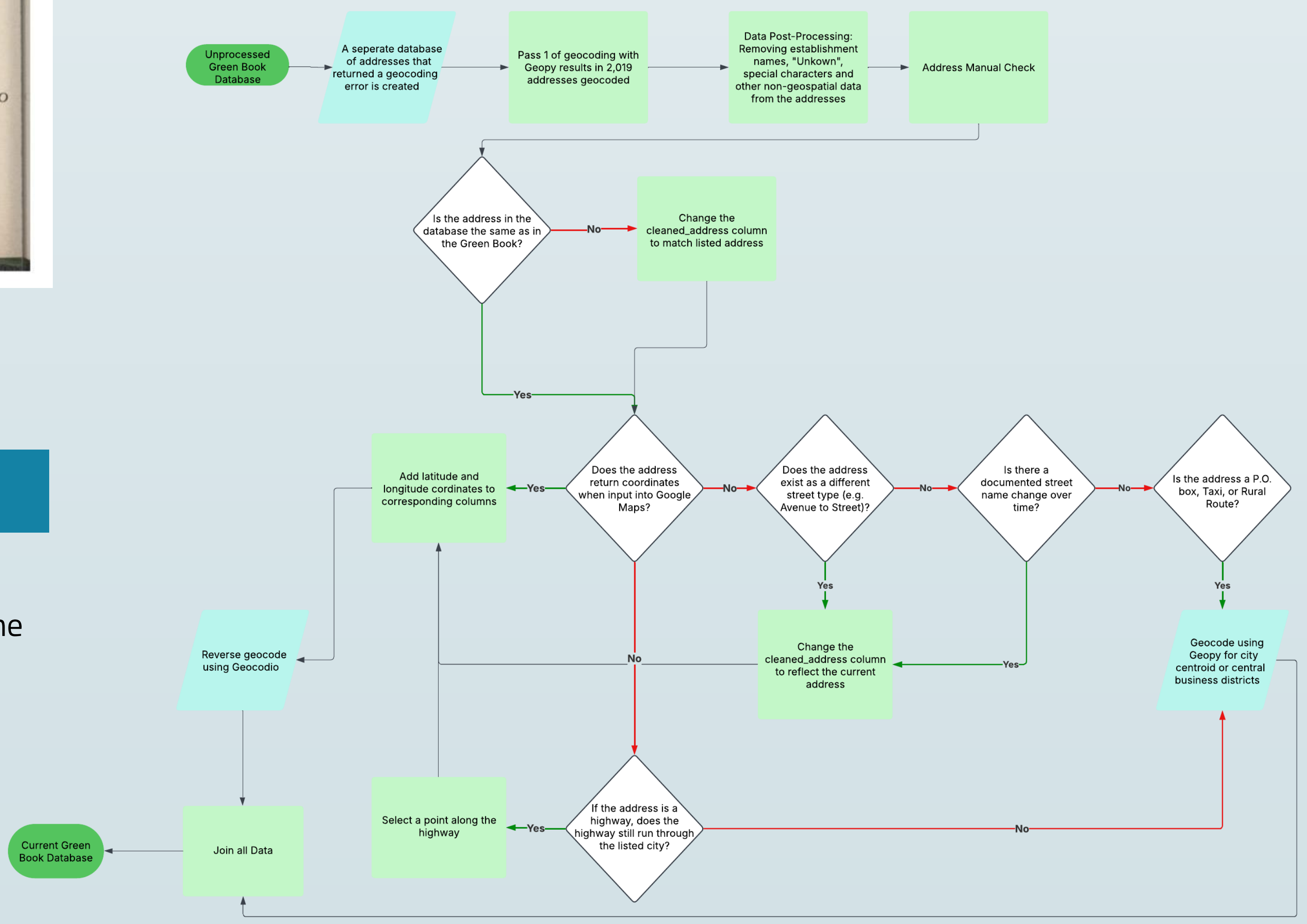

Unprocessed Green Book Database
A seperate database of addresses that returned a geocoding error is created
Pass 1 of geocoding with Geopy results in 2,019 addresses geocoded
Data Post-Processing: Removing establishment names, "Unkown", special characters and other non-geospatial data from the addresses
Address Manual Check
Is the address in the database the same as in the Green Book?
No
Change the cleaned_address column to match listed address
Yes
Does the address return coordinates when input into Google Maps?
Yes
Add latitude and longitude cordinates to corresponding columns
No
Does the address exist as a different street type (e.g. Avenue to Street)?
No
Is there a documented street name change over time?
No
Is the address a P.O. box, Taxi, or Rural Route?
Yes
Change the cleaned_address column to reflect the current address
Yes
Yes
Geocode using Geopy for city centroid or central business districts
No
Reverse geocode using Geocodio
If the address is a highway, does the highway still run through the listed city?
Yes
Select a point along the highway
No
Join all Data
Current Green Book Database

Figure 5. Generalized workflow of the geocoding process

Results

- A total of 39,935 addresses were extracted from Green Books across all years. Across all addresses extracted, 11.4% (N = 4,562) were unable to be forward geocoded due to preprocessing errors and Geocodio's lower tolerance level for incomplete or ambiguous addresses.
- Of the remaining addresses, 23.6% (N = 8,350) exceeded the threshold for address inaccuracy and 10.5% (N = 3,717) of addresses were extracted without a street name.
- A total of 44.3% (N = 2,019) of addresses with missing latitude and longitude coordinates were resolved via TIGER data. The remaining 55.7% (N = 2,543) of unresolved addresses revealed common errors including: incomplete addresses, changes in road type, digitization errors, and address changes over time.
- 66% (N = 1,692) of these addresses were manually resolved by updating AI extraction errors, spelling, and updating differences in road type. Of all the extracted addresses, 2.1% (N = 851) were geocoded by city due to the address no longer existing or incomplete address information (e.g., a P.O. Box listed).

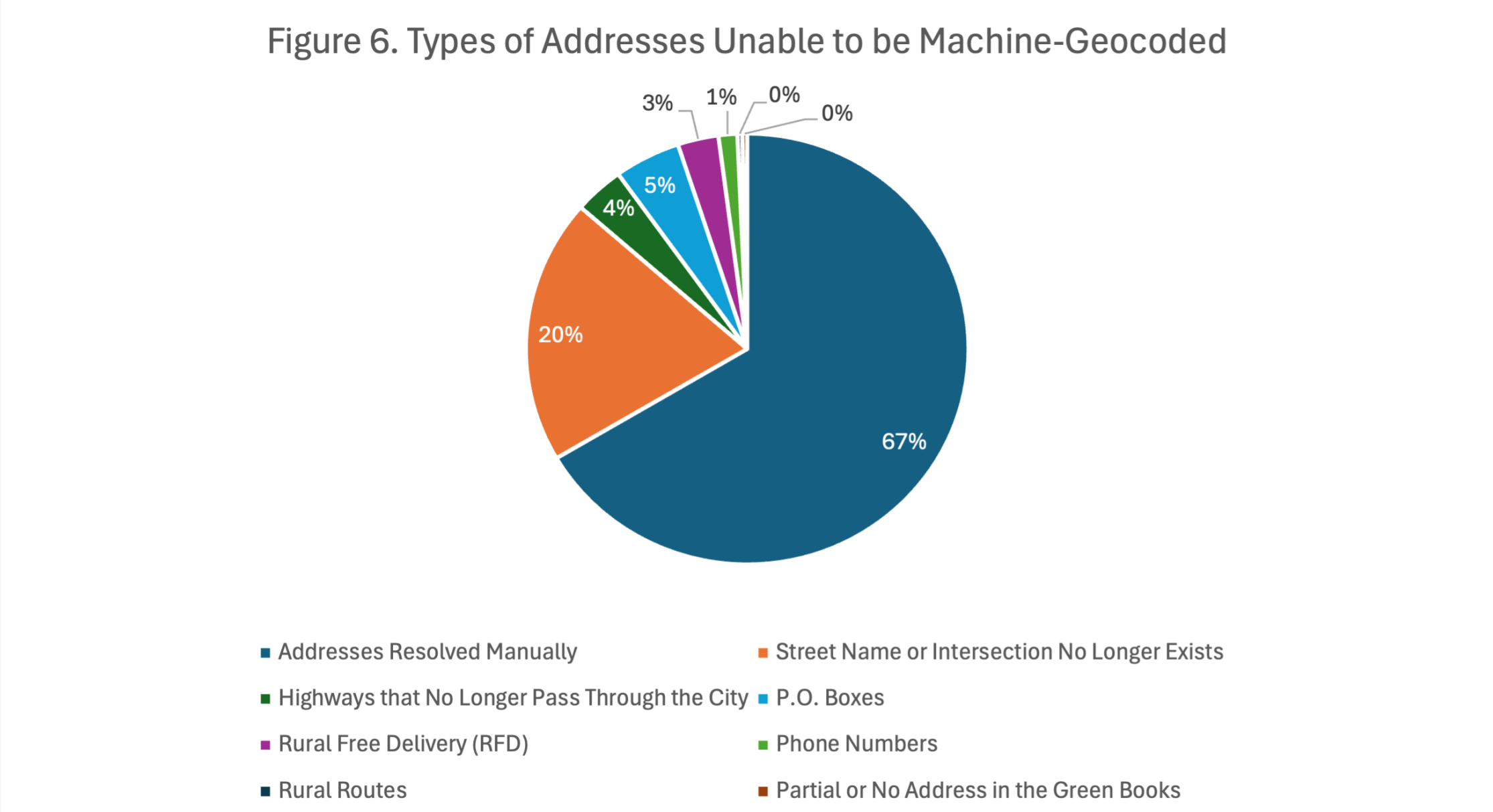


Figure 6. Types of Addresses Unable to be Machine-Geocoded

Figure 7. Number of Green Book establishments listed per year

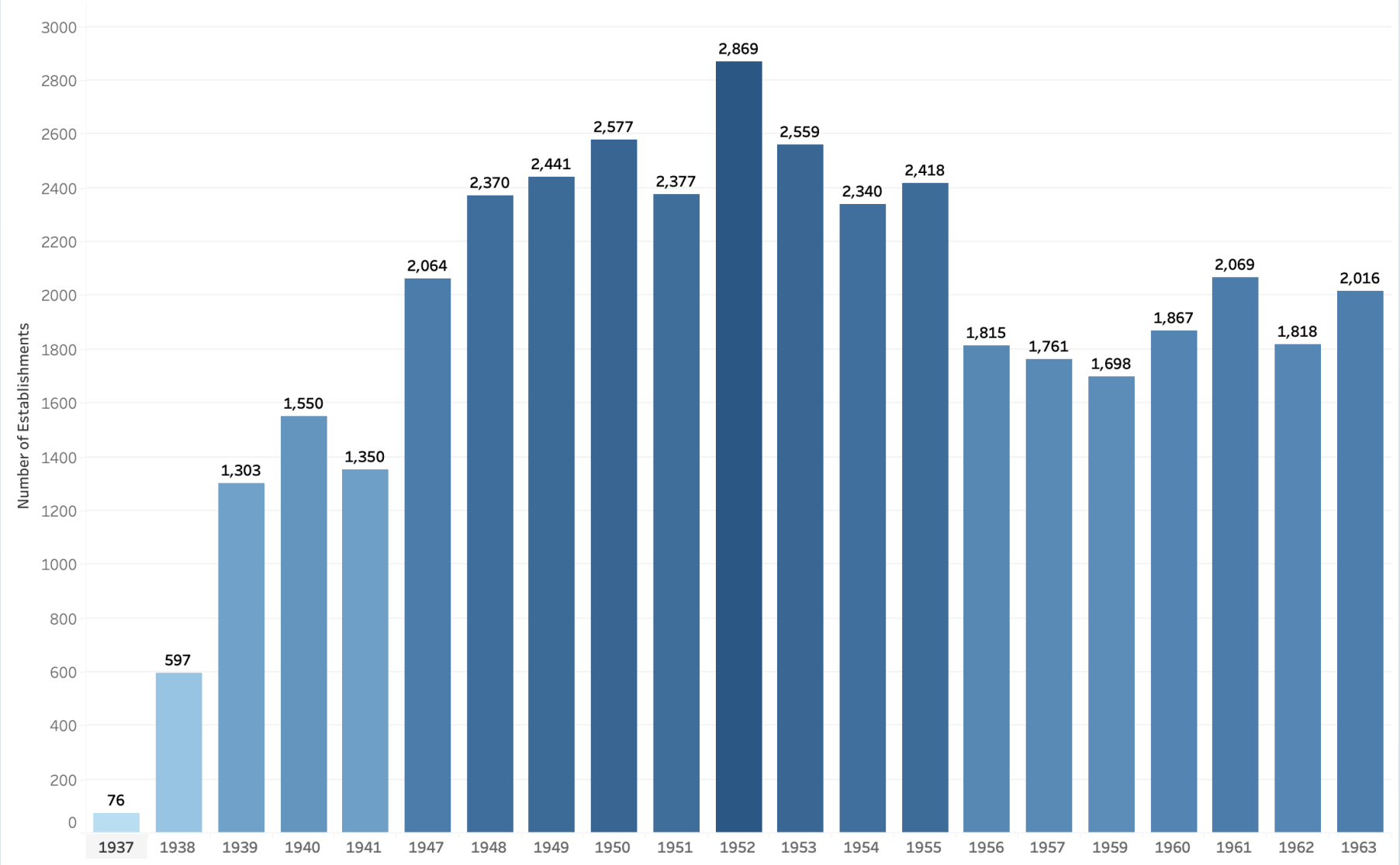
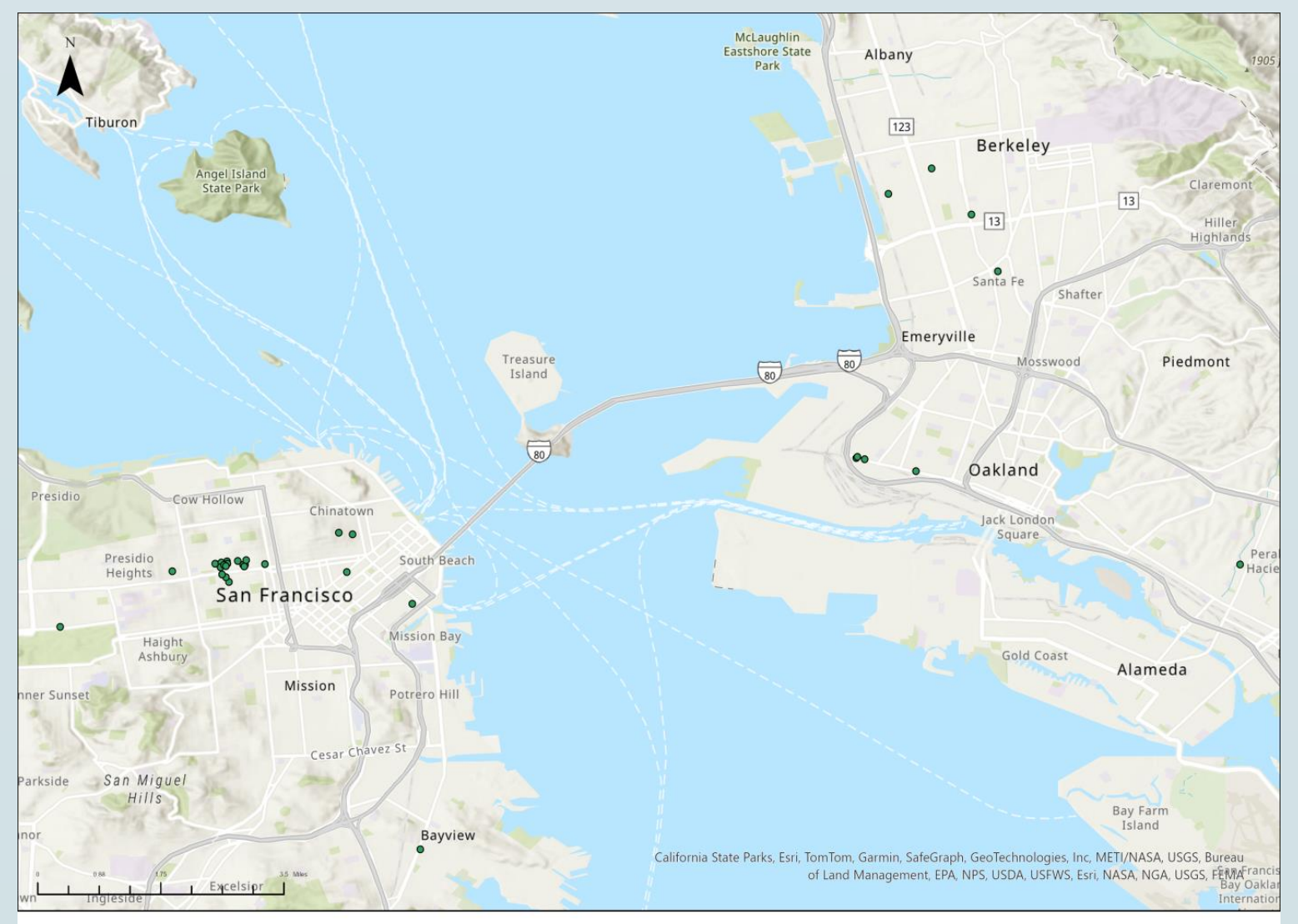


Figure 8. Map of Bay Area Green Book establishments (1950)



Discussion

Challenges:

- Errors in extracting city names due to page or column breaks, which resulted in "Unknown" or the establishment name being returned rather than the city.
- Special characters (¬Ω, Äô, Äù, Äù) were extracted in place of quotation marks and dashes in some addresses.
- Accounting for incomplete addresses, such as addresses containing P.O. boxes, taxicab phone numbers, and Rural Free Delivery (RFD).
- Working with a limited availability of block-group and block-level geospatial data for historical addresses to account for address changes over time.

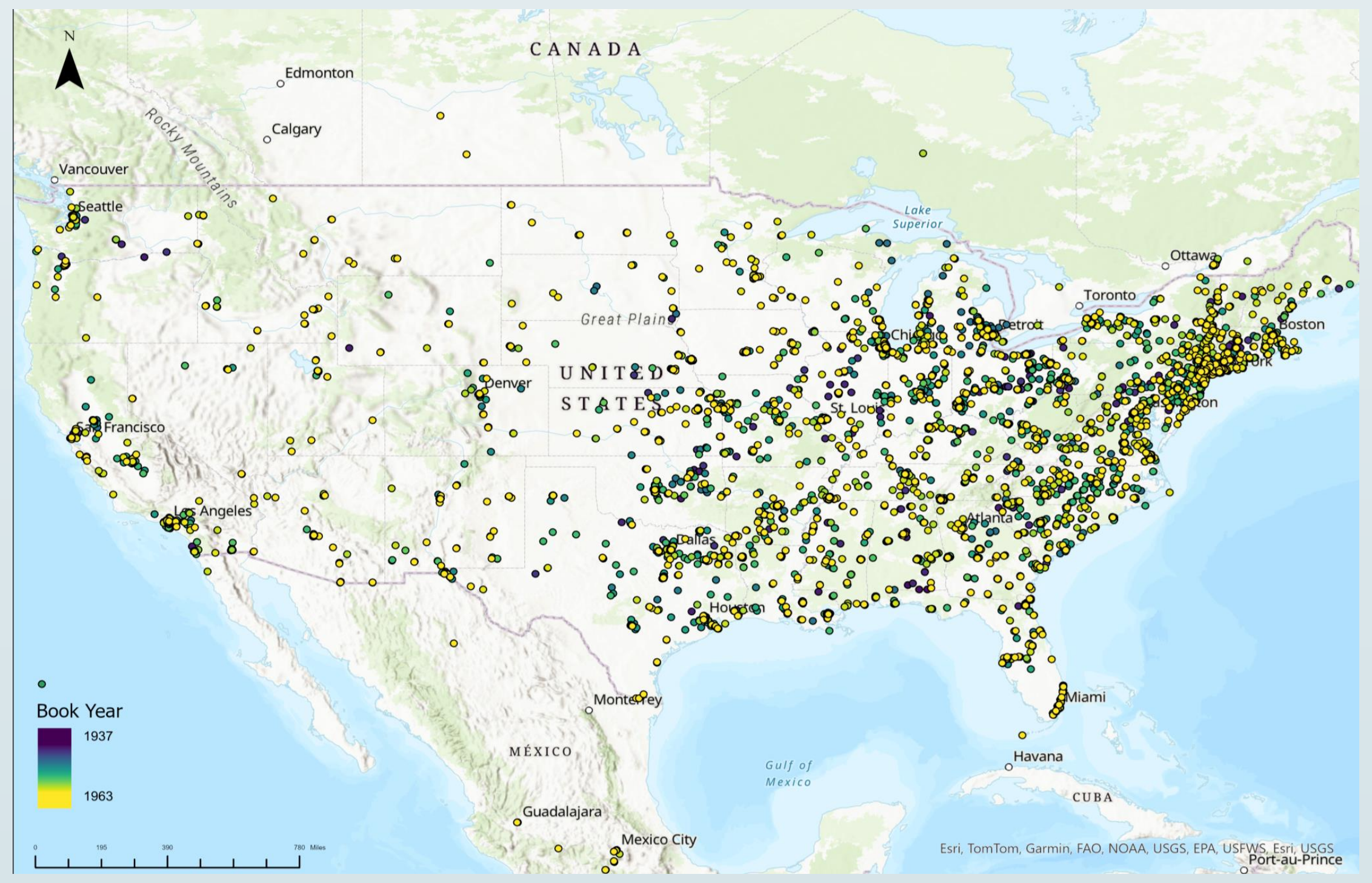
Next Steps:

- We plan on integrating business names and types to the database using few-shot learning to increase the consistency of model performance.
- Review addresses that failed fuzzy string matching to verify correctness and investigate discrepancies, which stem from the challenges listed above.

Applications and Future Work:

- Redlining and segregation associations with cognitive aging
- Combine with sundown town data to analyze stress exposure and terror on cognitive aging
- Food deserts and environmental hazards/pollution
- Economic mobility, urban renewal & gentrification/migration
- Loss of sense of place and community, time loss
- Access to transportation networks and commuting

Figure 9. Map of all Green Book establishments (1937-1963)



Conclusions

- This is a seemingly feasible, efficient application of AI that allows us to begin to explore more nuanced associations between aspects of structural racism and health.
- Our next steps include geospatial linkages to the cognitive aging data in the Health and Retirement Study.
- We will investigate how Black countermapping practices and access to Black spaces during the Jim Crow era, designed in response to structural racism, influence cognitive aging trajectories and risk of diseases of aging over time.

This work was supported by the National Institute on Aging Grant #: K23AG084871 (PI: Hill-Jarrett)