A Hierarchical Approach for Geocoding Birthplaces in Temporally Continuous Crowd-Sourced Family Tree Data

Maryam Torkashvand, Caglar Koylu

ABSTRACT: Geocoding is a fundamental yet complex step in temporal studies due to constantly evolving administrative borders and place names, and the uncertainty of geographic and temporal information. For example, identifying locations within crowdsourced datasets, such as family trees, is complex because recorded place names may be uncertain, inaccurate, and contain varying spellings (for instance, full names or abbreviations) and in inconsistent formats, such as mentioning only the country, or state, or a combination of city, county, state, and country. Moreover, place names and administrative boundaries drastically change over time, adding another layer of complexity to the geocoding process of fine-scale places. This paper presents a workflow for geocoding birthplaces of US-born individuals from crowd-sourced genealogical files spanning from 1789 to 1940. We introduce a method that geocodes the birth locations at the finest possible level by matching places with corresponding historical administrative boundaries within a range of individuals' birth years. Our preliminary study analyzing 72,335 trees with over 250 million individual records shows the potential of our approach for use in complex crowd-generated spatio-temporal datasets.

KEYWORDS: Geocoding, Spatio-temporal data, Historical data, Crowd-sourced, Place matching.

Introduction

Geocoding assigns latitude and longitude coordinates or administrative boundaries to addresses or place names by matching them with reference datasets and gazetteers. This process is a fundamental yet complex step in historical studies due to inconsistency in place names and changing boundaries over time. The centroid is often used for georeferencing regions, such as provinces and counties, which simplifies the entire area to a single point (Cura et al., 2018; Hedefalk et al., 2017). However, in cases with significant border changes over time, such as in the U.S. where historical borders have drastically changed, the actual location and extent of an administrative boundary become important. Several studies employ name directories and gazetteers, which may or may not include geometric data, to identify the location of historical places (Daras et al., 2015; Mertel et al., 2021; Walford, 2019). Berkes et al. (2023) used two U.S. historical databases to geocode residential locations from U.S. census microdata spanning 1790 to 1940 at sub-county levels such as town and small city. Similarly, polygons resulting from the overlay of contemporary and historical maps have been used for mapping and geocoding places in Canadian censuses at different scales (St-Hilaire et al., 2007).

Unlike commonly geocoded data at decennial intervals, geocoding crowdsourced data such as family trees with continuous event dates such as birthplaces and dates is complex because the recorded place names reflect the time periods in which they exist. Therefore, it's crucial to select the correct historical reference using available temporal information for accurate place matching (Hedefalk et al., 2018). Moreover, censuses capture national data, whereas family trees in the U.S. contain a mix of global and local birthplaces, creating challenges in filtering foreign locations due to inconsistent formatting. Geocoding family tree birthplaces, unlike standard geocoding with set spatial levels, deal with imprecise, non-uniform historical and crowd-sourced data, with locations ranging from countries and states to cities and townships. In this paper, we introduce a geocoding workflow for geocoding the birthplaces of individuals born in the U.S. between 1789 and 1930. We extract birth locations from 72,335 crowd-generated family trees from rootsweb.com, mainly from contributors in the U.S. and Canada (Koylu et al., 2021).

Method

Our method addresses the specific challenges of fine-scale geocoding of historical birth locations over a large temporal and geographical extent. Our workflow identifies and geocodes birthplaces, recorded in the GEDCOM files, to the finest available geographical scale (state, county, city, or township) using a hierarchical match. Figure 1 shows our geocoding workflow.

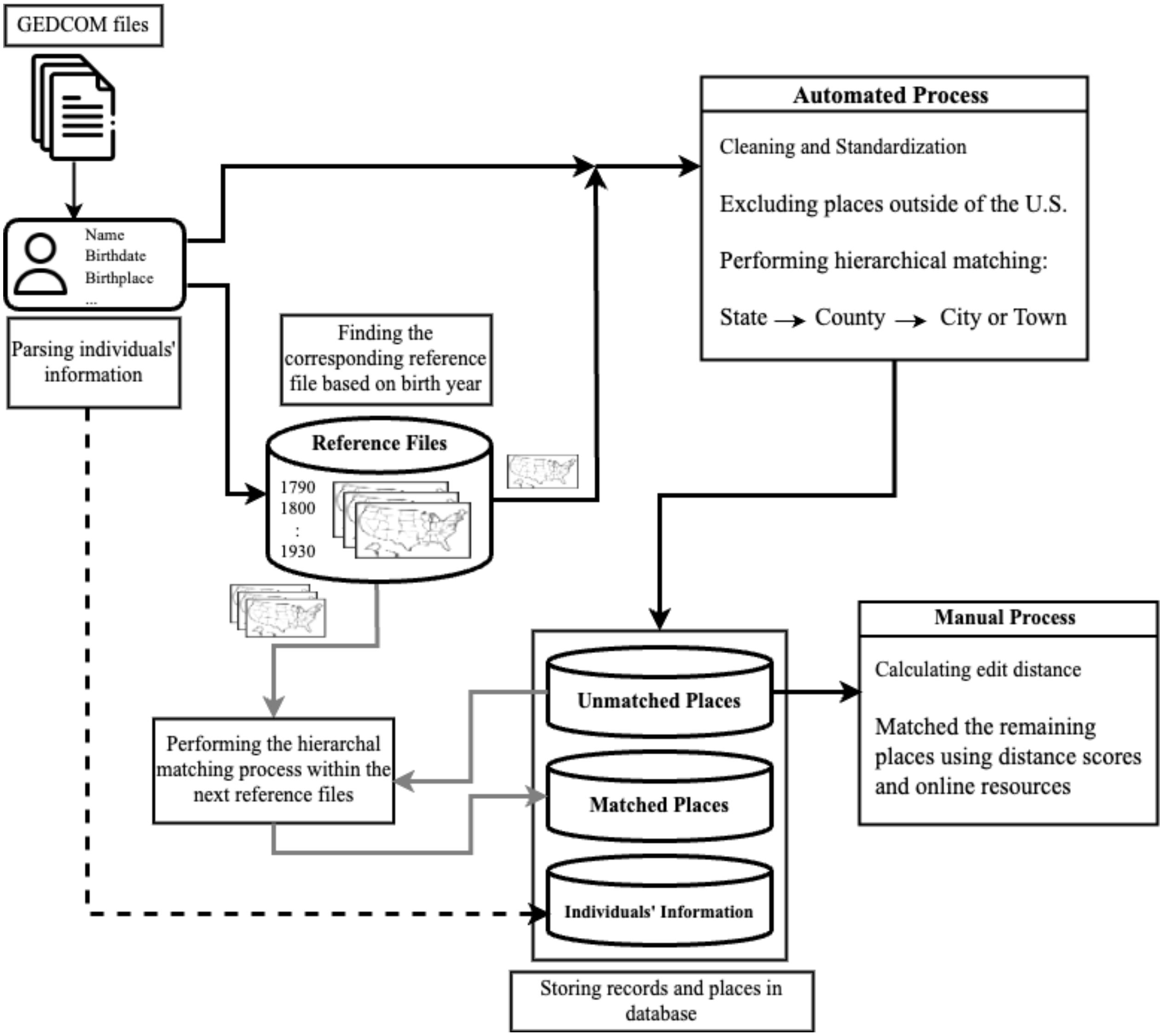

GEDCOM files
Name
Birthdate
Birthplace
...
Parsing individuals' information
Finding the corresponding reference file based on birth year
Reference Files
1790
1800
:
1930
Automated Process
Cleaning and Standardization
Excluding places outside of the U.S.
Performing hierarchical matching:
State → County → City or Town
Performing the hierarchal matching process within the next reference files
Unmatched Places
Matched Places
Individuals' Information
Storing records and places in database
Manual Process
Calculating edit distance
Matched the remaining places using distance scores and online resources

Figure 1. Workflow of Geocoding GEDCOM files.

Step 1. Parsing GEDCOM files

We begin by extracting birthplace and date information for each record in the GEDCOM files. In addition to the spatiotemporal data, we also parse and store other relevant fields, including individuals' names, parents' and spouse's names, and gender, for further analysis.

Step 2. Automatic place matching

We first standardize birth locations by removing punctuation and numbers, lowercasing, and stripping extra whitespaces. We also filter out locations outside the U.S. or those with insufficient detail. During the 19th century, as the United States expanded, there were significant changes in the boundaries and names of administrative units. We employ two collections as reference: township and city names from 1860 to 1930 censuses and county names and boundaries from IPUMS' National Historical Geographic Information System (NHGIS) for 1790-1860, both in decennial intervals. Based on available temporal snapshots, we select the correct decade to define accurate spatial boundaries for geocoding our dataset. We choose the closest next decade to the individual's birth year to determine the corresponding reference file for geocoding their birthplace. For instance, if someone was born between 1850 and 1860, we use the 1860 census data as the reference for the place match. If no match is found in the determined reference, we proceed to the next temporal reference, including earlier periods, and continue the process until we either find a match or exhaust all available reference files.

Next, we perform a hierarchical search beginning at broader administrative levels and progressively narrow down to more specific localities, with the finest geographic unit being a town or city, as identified in the GEDCOM files. The focus is on achieving the highest resolution of geospatial data accuracy by geocoding locations to the most precise geographic entity available. In GEDCOM files, birth locations typically follow a coarse-to-fine order, such as "*Van Buren, Crawford Co, Arkansas*". However, this is not always consistent. For example, the below line represents an entry for a birthplace:

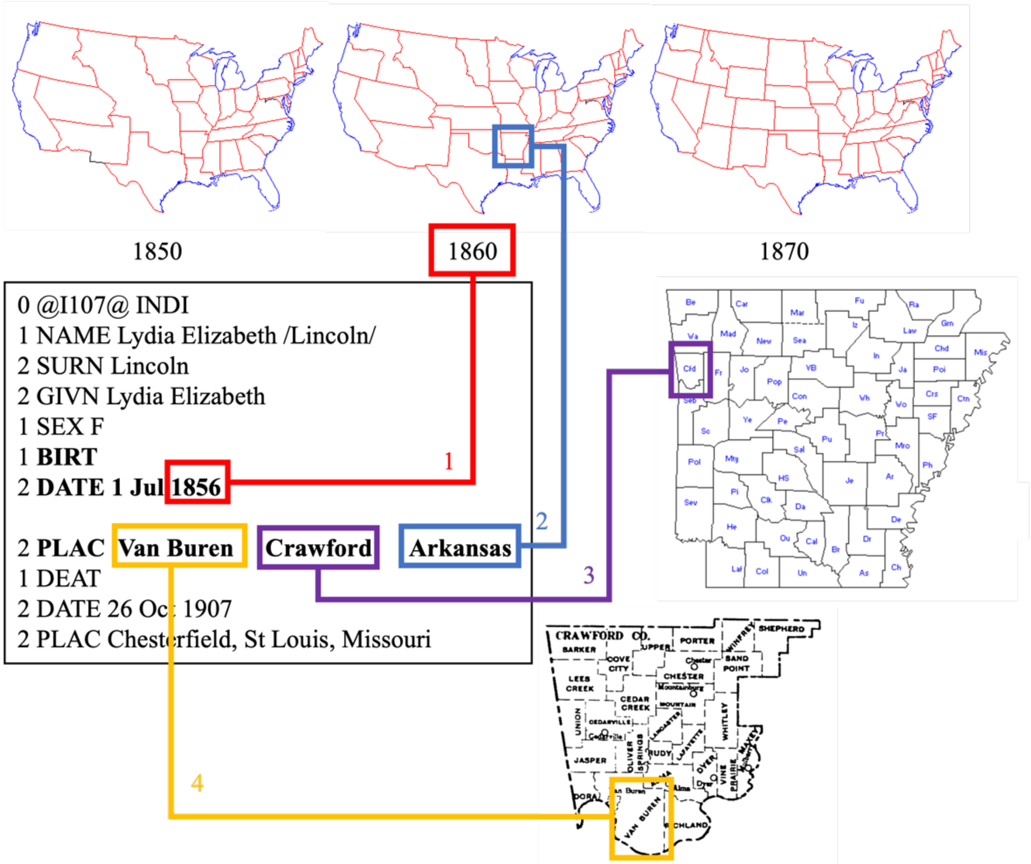
"he was born in Van Buren Crawford county in Arkansas USA".

Given the variability and lack of a consistent structure in place names, it is not always possible to directly split them into tokens based on a presumed fixed order of city, county, and state. To effectively handle this complexity, we implement the following procedure:

- We create a comprehensive dictionary of state names, including common abbreviations and misspellings.
- We search for state names within the birthplace string from right to left, using this dictionary. This approach helps handle cases where the county's name might also be a state name, such as *"texas county missouri".*
- Once the state name is identified, we search for county names within the found state. We use the county names from the appropriate reference table selected earlier to search for an exact match within the remaining part of the birthplace.
- After identifying the county, we look for city or town names within that county in the remaining part of the birthplace string (if it exists). If no county is identified, the birth

location might only mention a city, and we search for the cities within the found state. We can then use a 'spatial within' query to determine the county information of these cities.

Figure 2 illustrates a thematic presentation of the matching process, including the steps for determining the appropriate reference file and the hierarchical search process from state to county to city or town level for a given record. This figure shows a structured record with complete information to simplify the process.



```
0 @I107@ INDI
1 NAME Lydia Elizabeth /Lincoln/
2 SURN Lincoln
2 GIVN Lydia Elizabeth
1 SEX F
1 BIRT
2 DATE 1 Jul 1856

2 PLAC Van Buren  Crawford  Arkansas
1 DEAT
2 DATE 26 Oct 1907
2 PLAC Chesterfield, St Louis, Missouri
```

Figure 2. Hierarchical place matching process: 1. determine the corresponding decade for reference file based on birthplace, 2. search state names, 3. search county names within the found state, 4. Search city/township names within the found county.

Step 3. Further search and manual identification

Temporal historical reference files often lack comprehensiveness in detail and quality, especially those dates earlier than 1860. Furthermore, GEDCOM files are often created long after the actual birth years, and many users might enter the contemporary names of birthplaces, which may not exist around the birth date information. To improve the accuracy and match rates, we iteratively execute the search process step by step across all temporal reference files for unmatched places. Moreover, due to historical references often lacking city boundaries, we will address this gap by geocoding cities based on centroid points, using available historical references and online resources.

Finally, for birth locations that remain unmatched after the automated process, we calculate the edit distance (Levenshtein) to find the best possible matches. We apply a filter to select the top five places with the highest similarity score, using a matching threshold of 0.70. The human operator needs to search for the historical place names on online resources and collections of U.S. historical administrative boundaries and decide the final match for all unmatched birthplaces.

Preliminary Results

We processed 72,335 GEDCOM files containing spatiotemporal information for 250 million individuals born in North America and Europe between 1789 and 1930. Table 1 shows the success rates for matching locations at different geographic levels (state, county, city/township) during the automatic workflow phase. Initially, 50% of birthplaces were accurately geocoded at a sub-state level, with an 11% increase after expanding the search to other references. This improvement reflects the increasing quality of references over time, especially after 1860, with GEDCOM files created in the 20th and 21st centuries likely featuring more modern place names. Challenges remain with unmatched specific or misspelled locations outside the U.S., like "*rigaud vaudreuil québec*" or "*massachucetts*". Enhancements to the foreign location dictionary and misspelled names list are expected to significantly boost automated geocoding success.

Table 1: Initial match ratio of birthplaces from the automated phase

Total count of distinct strings from birthplace fields	*Ratio*					
	searching in corresponding reference				*searching in all references*	
	City/Township & County & State	County & State	Only State	Un-matched	City/Township & County & State	County & State
3,634,748	0.17	0.33	0.25	0.24	0.23	0.38

Conclusion

We introduced a workflow for geocoding temporally continuous crowd-sourced family tree data, which addresses locational and temporal uncertainties of crowd-sourced data. For future work, we will first link individuals to their geocoded birthplaces. Koylu et al. (2021) cleaned, deduplicated, and connected the same dataset to generate the largest connected family tree with 40 million individuals in a single pedigree. However, records in Koylu's population-scale tree are geocoded at the state level in the U.S. We will identify the matching individual from the cleaned records and retrieve their fine-scale geocoded birthplace from the original files. Moreover, using the state-level geocoded dataset as a benchmark, we will evaluate our method's accuracy by comparing fine-scale geocoded birthplaces and examining the consistency of birthplaces among siblings.

Acknowledgments

This work is supported by The National Science Foundation (NSF) Grant No. 2215568 titled "Population-scale kinship networks and migration". We extend our thanks to Tyler Watters for his assistance with this project and to Dr. Evan Roberts for providing the census reference files.

References

Berkes, E., Karger, E., & Nencka, P. (2023). The census place project: A method for geolocating unstructured place names. *Explorations in Economic History*, *87*, 101477.

Cura, R., Dumenieu, B., Abadie, N., Costes, B., Perret, J., & Gribaudi, M. (2018). Historical collaborative geocoding. *ISPRS International Journal of Geo-Information*, *7*(7), 262.

Daras, K., Feng, Z., & Dibben, C. (2015). HAG-GIS: A spatial framework for geocoding historical addresses. Proceedings of the 23rd GIS Research UK Conference, Leeds, UK,

Hedefalk, F., Pantazatou, K., Quaranta, L., & Harrie, L. (2018). Importance of the geocoding level for historical demographic analyses: A case study of rural parishes in Sweden, 1850–1914. *Spatial Demography*, *6*, 35-69.

Hedefalk, F., Svensson, P., & Harrie, L. (2017). Spatiotemporal historical datasets at micro-level for geocoded individuals in five Swedish parishes, 1813–1914. *Scientific data*, *4*(1), 1-13.

Koylu, C., Guo, D., Huang, Y., Kasakoff, A., & Grieve, J. (2021). Connecting family trees to construct a population-scale and longitudinal geo-social network for the US. *International Journal of Geographical Information Science*, *35*(12), 2380-2423.

Mertel, A., Zbíral, D., Stachoň, Z., & Hořínková, H. (2021). Historical geocoding assistant. *SoftwareX*, *14*, 100682.

St-Hilaire, M., Moldofsky, B., Richard, L., & Beaudry, M. (2007). Geocoding and mapping historical census data: The geographical component of the Canadian Century Research Infrastructure. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *40*(2), 76-91.

Walford, N. S. (2019). Bringing historical British Population Census records into the 21st century: A method for geocoding households and individuals at their early-20th-century addresses. *Population, Space and Place*, *25*(4), e2227.

Maryam Torkashvand, Ph.D. Student, Department of Geography and Sustainability Sciences, University of Iowa, Iowa City, IA

Caglar Koylu, Associated Professor, Department of Geography and Sustainability Sciences,

University of Iowa, Iowa City, IA