Geocoding Birthplaces in Temporally Continuous Crowd-Sourced Family Tree Data

Ariana Luan[1], Maryam Torkashvand[2], Chun Hang Chan[2] , Caglar Koylu, PhD[2],

[1]Amador Valley High School, [2]Department of Geographical and Sustainability Sciences, University of Iowa

Background

- Previous work by Koylu et al., 2021, generated, to date, the largest population-scale family tree, which connects 40 million relatives to their common ancestors using crowdsourced genealogical data (Koylu et al., 2020).
- Tree data contains information on names, birth and death dates and places, and kinship ties such as parent-child and spouse, which can be used to measure long term migration patterns and social processes in United States at state level (Koylu et al., 2022)

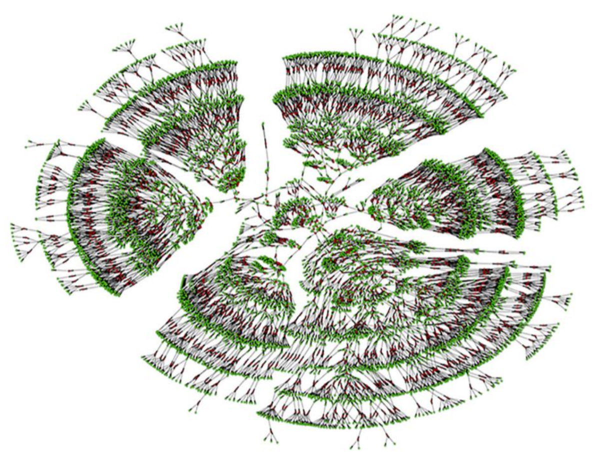

Figure 1 Example tree displaying around 6000 individuals (shown in green) and marriages (shown in red) over 7 generations. (Kaplanis et al., 2018)

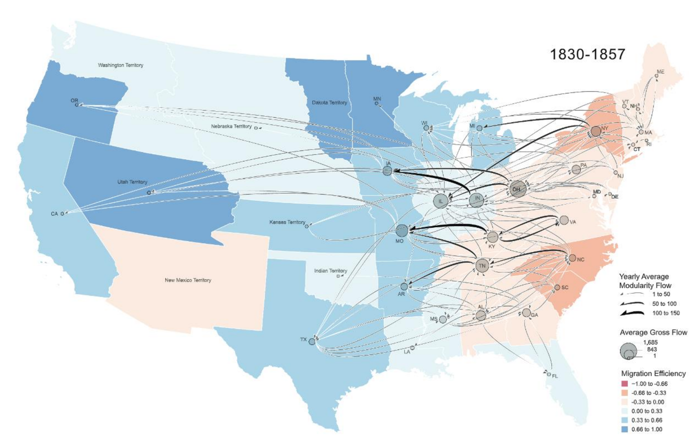

Figure 2 Example of a migration flow map derived from family tree data set. The choropleth map represents migration efficiency, flow lines represent the state-to-state migration flows (Koylu et al., 2022).

- Unlike census data, that only provide snapshots of population dynamics, tree data enables the study of a continuous evolution of population, movement, and social dynamics. However, this requires fine–scale geographic locations.
- In this research, we enrich the tree dataset by geocoding birthplaces within this data set from 1789 to 1940 on an even finer scale, down to **county or city level.**

Geocoding is the process of converting place names into geographic coordinates. However, geocoding historical places is challenging, especially for crowdsourced datasets. This works aims to address the challenges of historical geocoding:

- **Temporal factor** – constantly changing county and city names and boundaries
- **Crowdsourced data** – misspellings, inaccurate locations, inconsistent formatting

Results

We manually matched over 10,000 birthplaces. We found many recurring errors made for locations from both the AI-based and the code-based matching. Some of the most common errors were misrecognition of abbreviations, unnecessary phrases in the input flagged as city or county names, foreign locations marked as locations in the U.S., and confusion over administrative boundaries.

Examples of Common Errors in AI Matching:

- "stdbirthplace" contained the word probably: the "state" column would be filled out incorrectly, the city and county would be random places within the incorrect state, and the "exist" column would be "no" even if the state was within the United States.
  - Python script was written that first scanned for any state abbreviations, then string fuzzy-matched using the partial token set ratio for any shortened state names, and finally checked for the full state name to identify the correct state. Ninety-nine rows out of a sample of around forty thousand AI matches were flagged as having this issue.

task-637	probably in virginia	virginia		yes			task1_StateNocoNoci
task-738	probably nc	north carolina		yes			task1_StateNocoNoci
task-749	probably in kentucky	kentucky		yes			task1_StateNocoNoci
task-757	probably maryland	maryland		yes			task1_StateNocoNoci
task-361	probably in wales great britton			no			task2_StateNocoNoci
task-617	probably nj	new jersey		yes			task2_StateNocoNoci
task-35	probably md	maryland		yes			task3_StateNocoNoci
task-47	connecticut probably	connecticut		yes			task3_StateNocoNoci
task-455	probably tenn	tennessee		yes			task3_StateNocoNoci

- "stdbirthplace" contained a place name that identified the location specifically as a county (ex: included abbreviation "co" or word "county), the AI would also put a record in the "city" column if there was one present with the same name as the county

task-1	culpepper county virginia		virginia	culpeper	culpeper	yes
task-187	breckenridge co ky	kentucky	breckinridge		breckinridge	yes
task-592	orangeburgh co sc	south carolina	orangeburg	orangeburg		yes

Examples of Common Errors in Code-Based Matching:

- Five locations in the table below are all located in France, but have the phrase "la", which the code flagged as belonging to the state "Louisiana." This error was then manually corrected.

;la chapelle montlinard	la chapelle montlinard	la		1790		no
lac ;la biche	lac la biche	la		1880		no
;la cluse	la cluse	la		1790		no
;la durantaye	la durantaye	la		1790		no
;la durantaye bellechasse	la durantaye bellechasse	la		1790		no

- "stdbirthplace" input would contain the state "south dakota," the AI would mark the state as just "dakota." because of right-to-left string processing

aberdeen ;south dakota	aberdeen south dakota	dakota	1930
arlington kingsbury ;south dakota	arlington kingsbury south dakota	dakota	1910
badger kingsbury county ;south dakota	badger kingsbury county south dakota	dakota	1900
britton marshall county ;south dakota	britton marshall county south dakota	dakota	
brookings ;south dakota	brookings south dakota	dakota	

Conclusion and Future Directions

- We were able to identify over fifteen common issues within the AI matches and resolve five of them, so far. These issues were mostly due to the challenges of crowdsourced and unstructured data: inconsistent formatting, inaccurate and inconsistent place names, and other parsing problems.
- This work will allow for previously unmatched places to be accurately geocoded at the finest geographic scale as possible. Without the AI validation process, 48.09% of exact places were able to be matched. In the future, the hope is that 70 or even 80% of locations will be accurately geocoded.
- Fine-scale geocoded locations opens doors to tracking families through space and time and analyzing broader population dynamics, such as migration patterns and kinship networks.

Methods

- We introduce an AI-based historical geocoding workflow to address the uncertainties in geocoding of historical crowdsourced data.
- Genealogical data stored in GEDCOM file format: name, birth date, birthplace, mother's name, father's name, etc.
- The closest possible census to birth date is found to match birth location within the correct administrative boundaries
- Hierarchical matching performed to identify first the state, then county, then city/township.

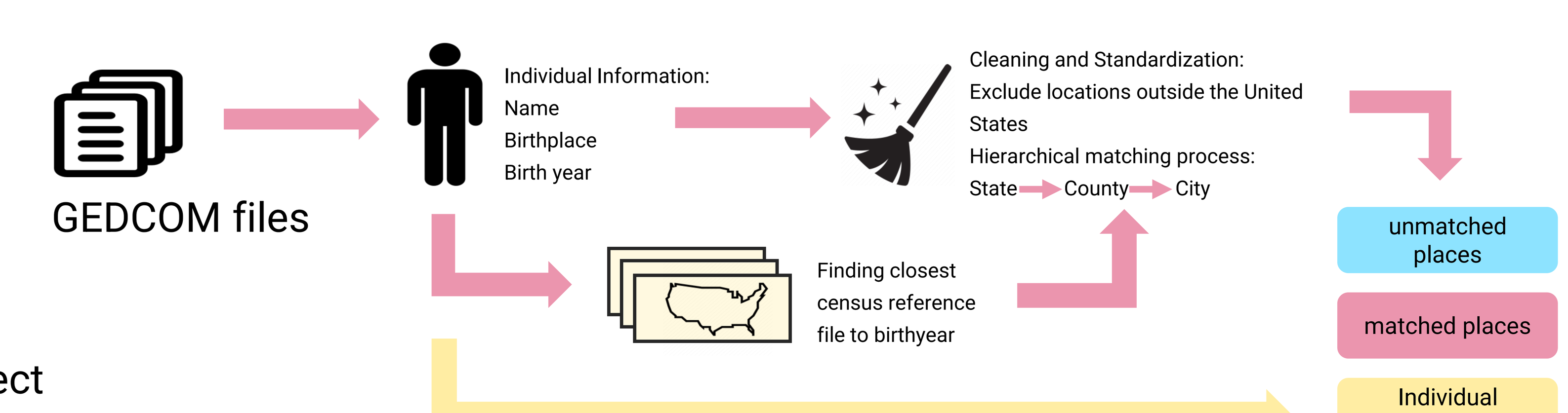

Figure 3 Overall process of geocoding. Individual information is extracted from the GEDCOM files, matched to the closest census reference (census taken in decennial intervals), and the location is matched hierarchically, from state to county to city level. Unmatched places, matched places, and individual information are then stored for further use and analysis.

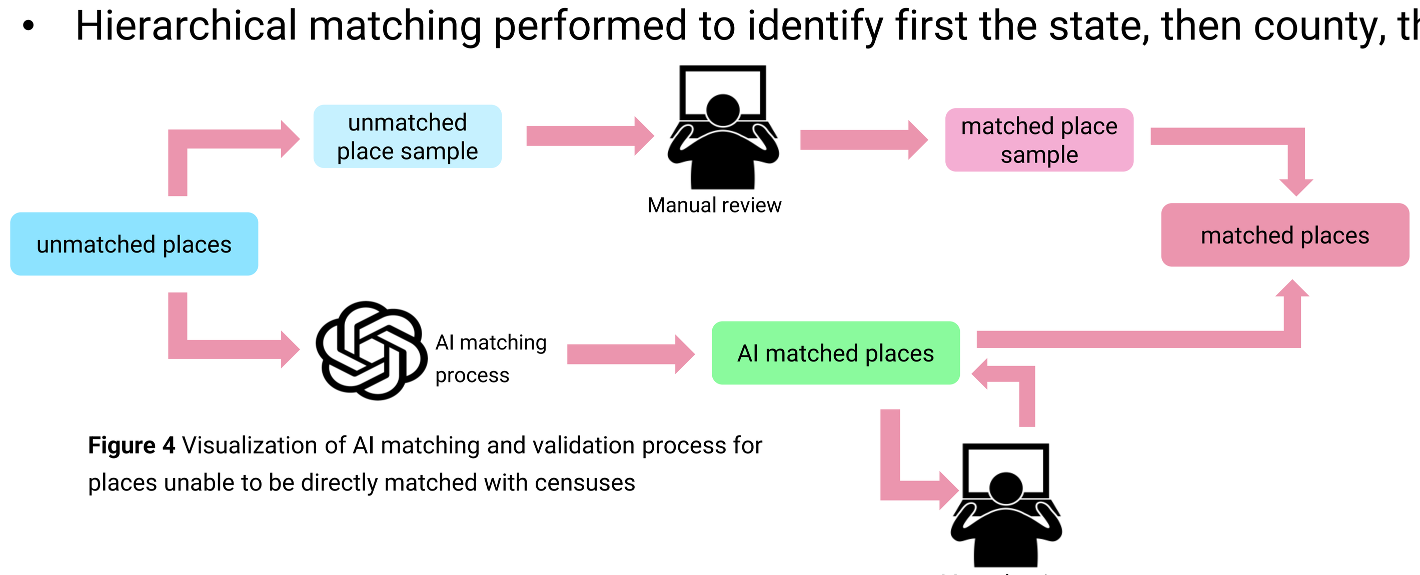

Figure 4 Visualization of AI matching and validation process for places unable to be directly matched with censuses

- Locations unable to be located directly within census references are attempted to be matched by AI

AI Validation Component

- Two main processes conducted: manual matching of unmatched places and validation and cleaning of large language model matches
- Recognize recurring problems and issues from matches with and without AI assistance
- Write scripts to pull them into separate CSVs and correct them into the proper format for further use

Acknowledgements

I want to thank all of my mentors in the GeoSocial lab: Dr. Koylu, Maryam, Loretta, and Henry. Thank you all for being so supportive and helpful throughout this process! I also want to recognize my lab mates Devesh and Sumin for being so encouraging, our collaboration was invaluable. Thank you to the Belin Blank Center for granting me this research opportunity. Finally, I want to thank of my friends here, the people I've met are just as important as the research I've done!

Full List of References

Selected References

1. Kaplanis, J., GordoSheikh, M., Gymrek, M., Bhatia, G., MacArthur, D. G., Price, A. L., & Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, *360*(6385), 171–175. https://doi.org/10.1126/science.aam9309
2. Koylu, C., & Bee Kasakoff, A. (2024). Population-Scale kinship networks. *International Encyclopedia of Geography*, 1–12. https://doi.org/10.1002/9781118786352.wbieg2193
3. Koylu, C., & Kasakoff, A. (2022). Measuring and mapping long-term changes in migration flows using population-scale family tree data. *Cartography and Geographic Information Science*, *49*(2), 154–170. https://doi.org/10.1080/15230406.2021.2011419
4. Koylu, C., Guo, D., Huang, Y., Kasakoff, A., & Grieve, J. (2020). Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S. *International Journal of Geographical Information Science*, *35*(12), 2380–2423. https://doi.org/10.1080/13658816.2020.1821885
5. Torkashvand, M. (2024). A Hierarchical Approach for Geocoding Birthplaces in Temporally Continuous Crowd-Sourced Family Tree Data. CAGIS+ UCGIS Symposium 2024