https://doi.org/10.14428/rete.v14i0/Locations

Geolocalization and the birth-to-death distance

David de la Croix
IRES/LIDAM, UCLouvain

Rossana Scebba
IRES/LIDAM, UCLouvain &
Early Modern History, KU Leuven

This note compares our experiences of manual vs programmatic geolocalization. As an application, we provide some results on the distance between place of birth and place of death of the RETE scholars, and showcase a map of RETE's locations as birth sources and death attractors.

1 Manual vs programmatic geolocalization

Since the start of the RETE project in 2017, contributors to the database were asked to georeference all locations manually, including longitude, latitude and current country. These locations include places of birth, activity (university or academy), and death. This encoding was based on searching Wikipedia for place name, and reporting the given longitude and latitude. Sometimes, Wikipedia had no entry about a place, and a search on Google Maps was performed, often yielding a reasonable result. Now, in January 2025, 13093 locations were found using these methods. It is now time to compare these results with what could be found using a structured and partially automated geolocalization procedure.

Geographic Information Retrieval (GIR), toponym resolution, place name linkage or disambiguation, and geo-labeling are various terms describing processes that associate geographic references with correct entities. This typically requires two elements: a list of locations and an authoritative dataset. In our case, we cross-check the RETE locations against Wikidata, a structured knowledge base known for its reliability in resolving ambiguities and linking geographic entities, even when names are ambiguous or contextually sparse (Hu, Janowicz, and Prasad 2014).

Using a Python pipeline, we queried the Wikidata API to retrieve unique Wikidata IDs for location names and fetched geographic coordinates (latitude and longitude) via the P625 property. The results were integrated into the original dataset, with new columns storing the Wikidata ID, latitude, and longitude for each location. To validate the accuracy of the assignments, we computed the distance between the original coordinates and Wikidata-derived ones. Locations with discrepancies greater than 5 km or missing matches (around 2,500 entries) were flagged and manually reviewed using OpenRefine (Delpeuch et al. 2024), a tool with an interface and a reconciliation feature that connects directly to Wikidata for precise matching.

The automated procedure, combined with the manual reconciliation, allowed to allocate 12306 locations to a Wikidata ID. This represents 94% of the manually encoded locations. Among the 787 locations which could not be found automatically, 683 suffer from ambiguity problem, while 69 were just not found in Wikidata. The 35 remaining locations had other issues such as wrong country allocation and duplicates.

Considering the 12306 locations found on Wikidata, we can compare the latitude and longitude found manually with the coordinates on Wikidata. We found that 469 locations have a gap larger than 10 kilometers. These discrepancies arise from multiple reasons. First mistakes in manual encoding (like the longitude was incorrectly recorded with a minus sign). Second, countries or provinces, for which Wikidata gives the coordinates of the capital while other sources give the coordinates of the centroid. Third, mispelling in the sources concerning the scholars. Four, wrong match in Wikidata.

On the whole, it seems we cannot completely avoid manual controls. However, there is a key difference compared to the initial round of manual searches: this time, we were more precise. We addressed ambiguities systematically and avoided approximations. This was facilitated by comparing the results against an existing thesaurus like Wikidata, which is interconnected with Wikipedia, rich of disambiguation pages for several entities. Still, automatically matching locations to Wikidata greatly enhanced scalability by limiting manual verification to less than a quarter of the full location dataset.

2 Distance between place of birth and place of death

Precise geolocalization can be used to study human mobility. One way to measure the mobility pattern of a group of people over time is to look at the distribution of the distance between place of birth and place of death. We compute such distance for each scholar using the Haversine formula,[1] before comparing our results with what is found in the literature.

For example, Serafinelli and Tabellini (2022) look at a sample of "creative people" from Wikipedia. They analyse the distribution of the distance for different periods, expecting to find an increasing mobility of time thanks to the consolidation of states and the improvements in the means of transportation throughout centuries. Their Figure 6 displays the distribution of birth-to-death distances in 1000, 1300, 1600, and 1800. The distribution remained surprisingly quite stable over time.

Spencer and Otterstrom (2024) compute the cumulative distributions of 2.83 million North American birth-death distances, from 1620 to 1980. Here too the cumulative distribution is stable until 1710, with a percentage of stayers around 40% (those who died close (< 1km) to their place of birth). Then, following the expansion of the Atlantic trade (Acemoglu, Johnson, and Robinson 2005), more people are movers. Median distance is not reported, but from the graph, it is slightly above 10km before 1710, and around 100km after.

Combining data from Freebase.com, the General Artist Lexicon, and the Getty Union List of Artist Names (ULAN), Schich et al. (2014) analyze over 150,000 individuals across two millennia to study aggregate birth-to-death migration across two millennia. The distribution of birth-to-death distances for the notable individuals in their sample from before 1300 CE to 2012 CE is fat-tailed, meaning most individuals migrated short distances, while only a few undertook long-range migrations. The median migration distance remained relatively stable over time, ranging from 214 km in the 14th century to 382 km in the 21st century, with a low of 135 km in the 17th century. According to the authors, this modest increase suggests that, despite global interconnectedness, most individuals historically remained within regional bounds. However, the tail of the distribution expanded significantly, reflecting the rise of long-range mobility driven by factors such as global colonization and increased transcontinental traffic.

Figure 1 shows the cumulative distribution of birth-to-death distance for the RETE scholars. It is based on 27985 scholars for whom places of birth and death are known (excluding those for which only countries of birth are known). The percent of "stayers" appears as the vertical axis intercept. It is equal to 26.4%. The median distance is 98km.

Let us now divide our sample in three: Middle-ages (<1492), Age of Discovery (1492-1699), Age of Atlantic Commerce (1700-1800). The sample sizes are respectively 2249, 11661, 14075. The percentage of stayers is, respectively, 32.3%, 24.2% and 27.1%. The median distance is 77km, 111km, 90km. The three cumulative distributions are relatively similar, echoing the result of Serafinelli and Tabellini (2022). Still we find that mobility increased from the Middle Ages to the Age of Discovery. The Age of Atlantic Commerce did not see any further increase in mobility, and, if anything, slight drop in the distances. This contradicts the finding of Spencer and Otterstrom (2024) for the whole US population.

[1] The haversine formula was used to calculate great-circle distances between geographic coordinates, as described by Sinnott (1984). The implementation follows the standard formula for computing distances on a sphere, using latitude and longitude in radians.

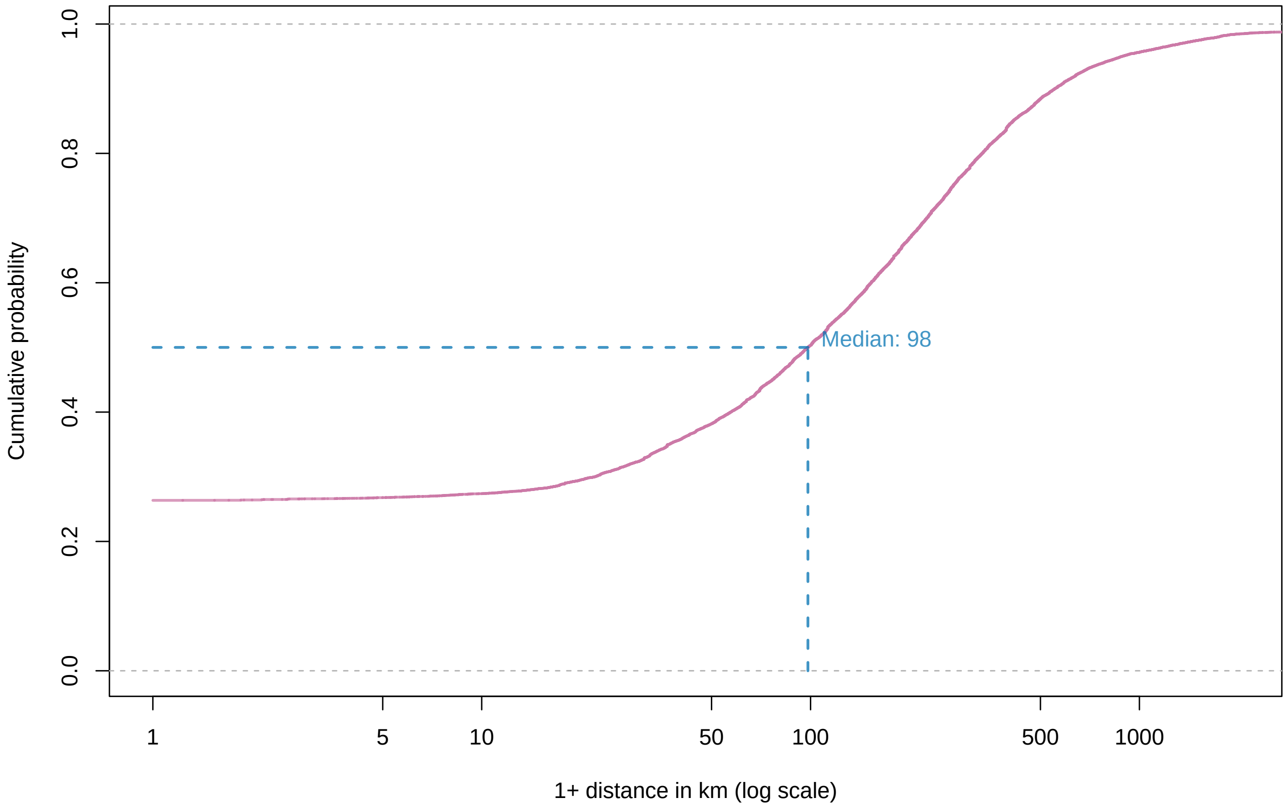

Figure 1: Cumulative distribution of birth-to-death distance

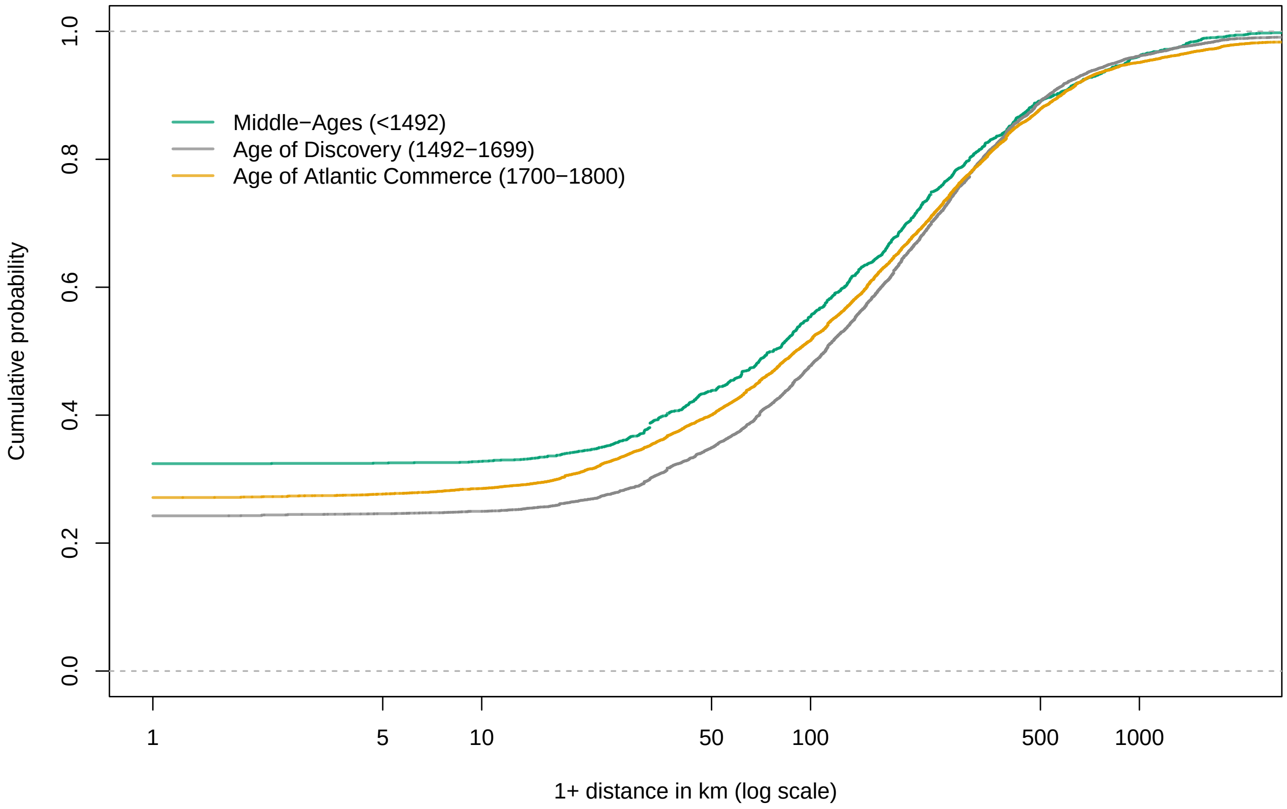

Figure 2: Cumulative distribution of birth-to-death distance by period

In Figure 3, we isolate scholars with an affiliation to the Universitas Lovaniensis (1425–1797, see Catoire et al. (2021)) – which celebrates its 600th anniversary on December 9 this year– from the rest of the RETE population. While the median distances are similar between the two groups, the distribution reveals differences in the lower and upper quartiles. Louvain scholars are less represented among those with very short birth-to-death distances (Q1) but become more concentrated starting from around 25 km, particularly within the 100–500 km range (approximately Q3), indicating a tendency toward shorter distances overall.

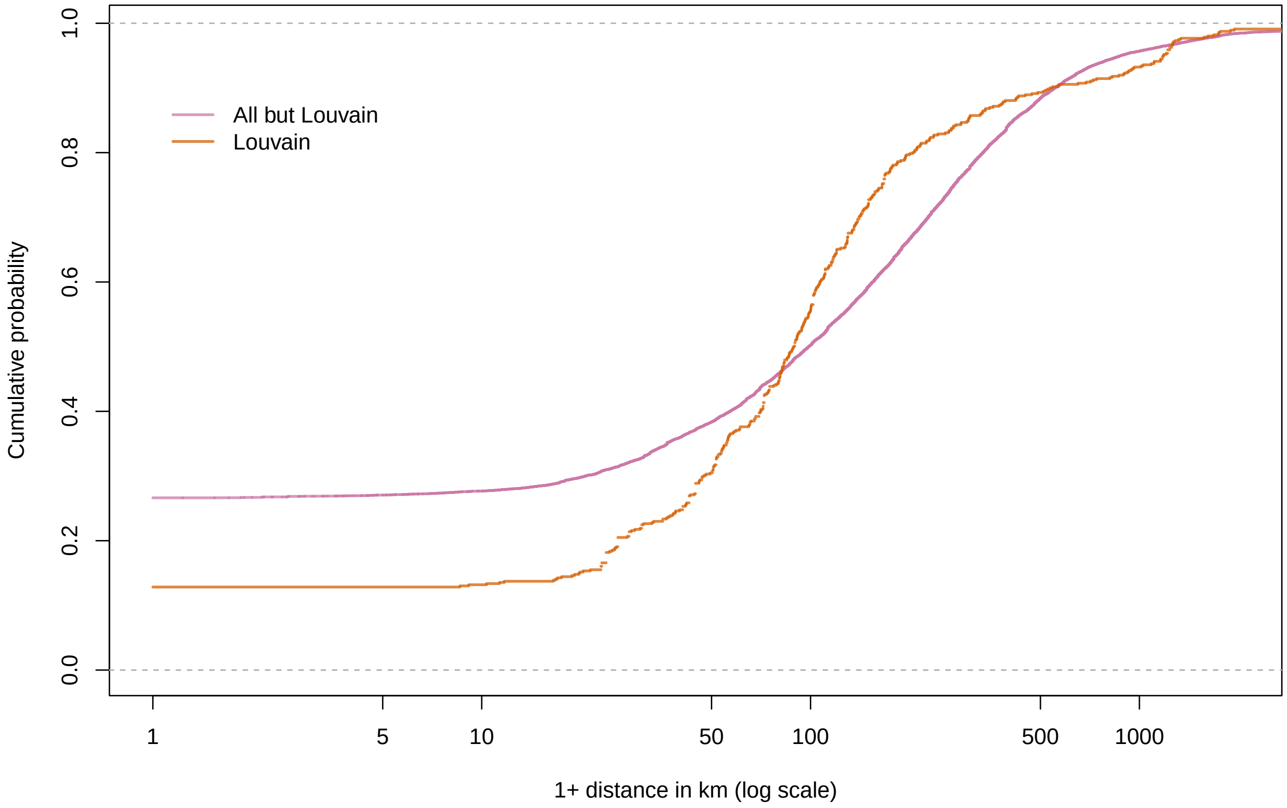

Figure 3: Cumulative distribution of birth-to-death distance by period, comparing scholars affiliated with the Universitas Lovaniensis (1425–1797) to all other scholars.

Finally, to analyze cultural mobility and the prominence of historical locations, we replicate Figure 1E from Schich et al. (2014) by mapping the raw counts of scholars' birth and death places in Figure 4. Consistent with the findings in Schich et al. (2014), we observe that European capitals, such as Paris and Rome, being historically highly important cultural hubs, serve as major death attractors. Central Europe also emerges as a region of moderate intensity for death attraction, due to its decentralized cultural landscape with multiple competing centers. Compared to Figure 1E from Schich et al. (2014), however, our map reveals a slight overrepresentation of Italian cities as birth sources, largely due to the prevalence of toponymic identifiers associating obscure Northern Italian scholars with their birthplace but often lacking corresponding death place records.

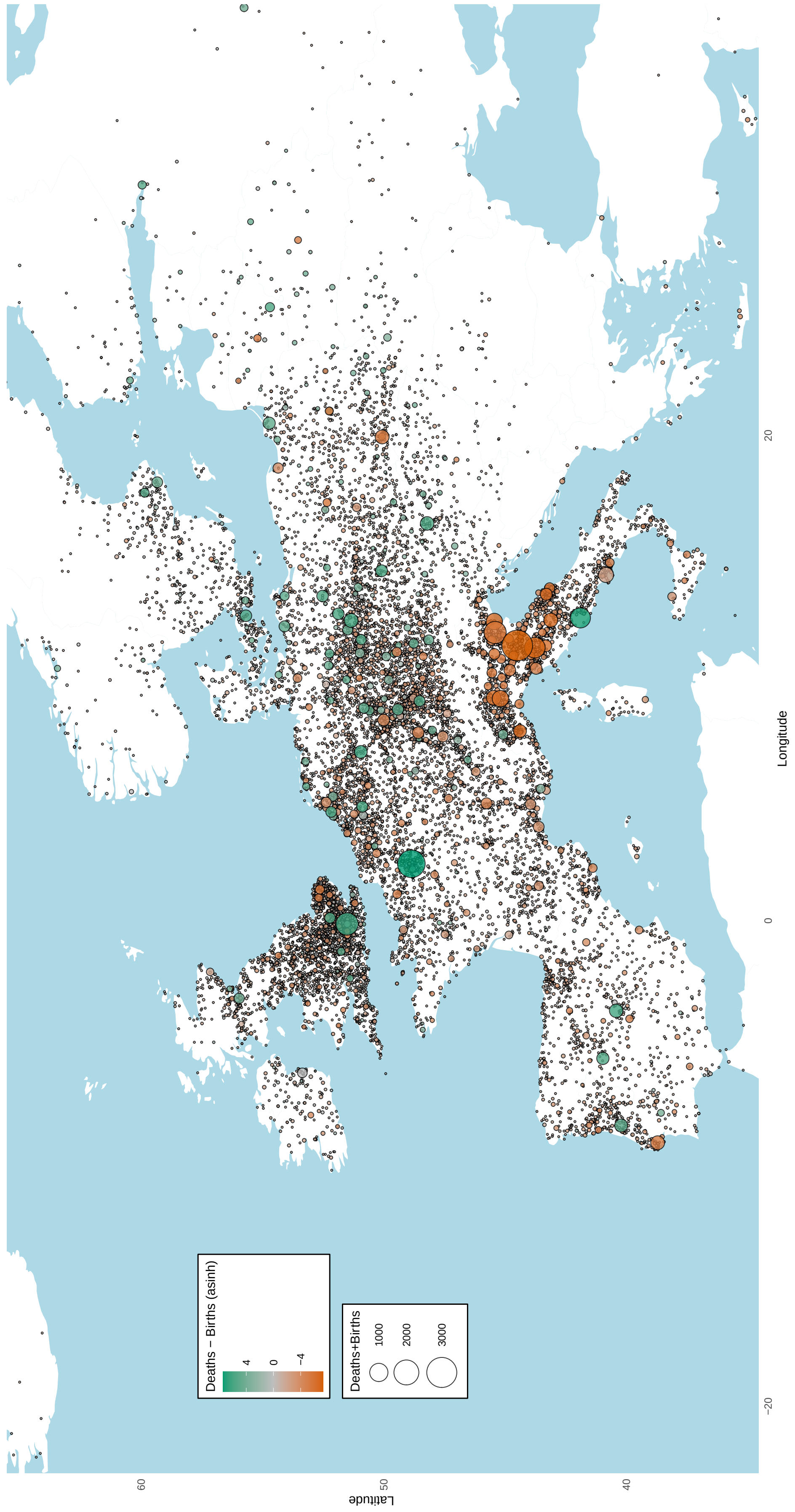

Figure 4: Map of cities as birth sources and death attractors.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 883033 "Did elite human capital trigger the rise of the West? Insights from a new database of European scholars."

Homepage: https://perso.uclouvain.be/david.delacroix/uthc.html
Database: https://shiny-lidam.sipr.ucl.ac.be/scholars/

First version 20 January, 2025.

References

Acemoglu, Daron, Simon Johnson, and James Robinson. 2005. The rise of europe: atlantic trade, institutional change, and economic growth. *The American Economic Review* 95 (3): 546–579.

Catoire, Guillaume, Valentine Debois, David de la Croix, Julie Duchêne, Maximilian Ganterer, and Mara Vitale. 2021. Scholars and Literati at the Universitas Lovaniensis (1425–1797). *Repertorium Eruditorum Totius Europae* 4:53–66. https://doi.org/10.14428/rete.v4i0/louvain.

Delpeuch, Antonin, Tom Morris, David Huynh, Weblate (bot), Stefano Mazzocchi, Jacky, Thad Guidry, et al. 2024. *Openrefine/openrefine: openrefine v3.8-beta1,* February. https://doi.org/10.5281/zenodo.10689569.

Hu, Yingjie, Krzysztof Janowicz, and Sathya Prasad. 2014. Improving wikipedia-based place name disambiguation in short texts using structured data from dbpedia. In *Proceedings of the 8th workshop on geographic information retrieval,* 1–8.

Schich, Maximilian, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. 2014. A network framework of cultural history. *Science* 345 (6196): 558–562.

Serafinelli, Michel, and Guido Tabellini. 2022. Creativity over time and space: a historical analysis of european cities. *Journal of Economic Growth* 27 (1): 1–43. https://doi.org/10.1007/s10887-021-09199-6.

Sinnott, Roger W. 1984. Virtues of the Haversine. *Sky and telescope* 68 (2): 158.

Spencer, Robin W, and Samuel M Otterstrom. 2024. Massive genealogies distinguish frontier from steady-state internal migration. *arXiv preprint arXiv:2410.18235.*