Geocoding Cities in the US Census, 1880-1950

Johann Ohler with
London School of Economics

James Feigenbaum
Boston University

WIP

April 7th

Motivation

Figure 1: Westwood, circa **1900**.

Figure 2: Westwood, circa **1950**.

- 1880-1950 was an important phase of urban development in the US (Kim & Margo 2003):
  - Migration (e.g. Boustan 2010), Residential Segregation (e.g. Cutler et al. 1999), Redlining (e.g. Fishback et al . 2023)

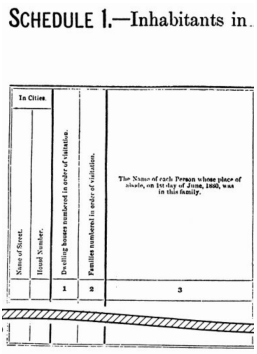
Motivation: Increasing Granularity

- Full-count US census (pre-1950) has ***rich* individual-level data** but ***poor* geographic granularity**.
  - Smallest geolocated identifiers at the county or city level.
  - Census Places Project remedies this problem for rural places (Berkes, Karger, and Nencka 2023).

What does granularity buy us?

- *Identifying variation*: ~3,000 counties vs. ~150,000 enumeration districts.
- *Longitudinal within-city data*: LR neighbourhood panel data.

Who lived where (in cities)?


SCHEDULE 1.—Inhabitants in

In Cities

Name of Street.

House Number.

Dwelling houses numbered in order of visitation.

Families numbered in order of visitation.

The Name of each Person whose place of abode, on 1st day of June, 1880, was in this family.

1 2 3

Figure 3: Census Enumeration Form

What we do

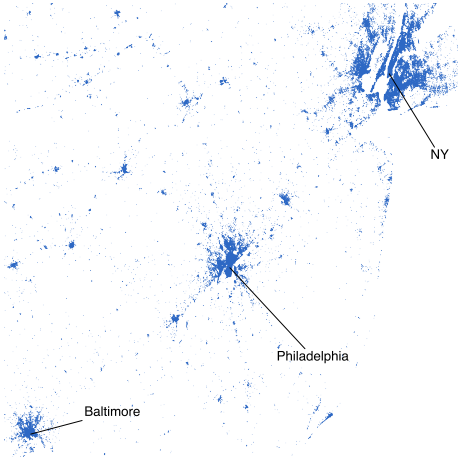


Figure 4: All geocoded EDs in 1950.

- We geocode all urban (and some rural) enumeration districts (EDs).
  - Smallest consistently available geographic identifier (1880-1950);
  - ~1500–3000 inhabitants.
- Leverage street intersections to determine ED location.
  - Match historical intersections to contemporary coordinates from OSM;
  - Compute centroid per ED.
- Crosswalk from ED ID to centroid coordinate.
  - Approximate within-city HH location (±250 m).
  - Project data onto consistent neighbourhood panel.

Contributions

1. Introduce a generalisable framework for geolocating spatial units in the absence of historical shape files.
   - Approach is transferable to other census w/ street/place information.

2. Improving the *usability* and *accessibility* of the US full-count census data:
   - Linking individuals across Census waves: Census Linking Project (Abramitzky et al. 2025), Census Tree (Buckles et al. 2024), ML Linking (Feigenbaum 2016).
   - Identifying granular geographic locations: Census Places Project (Berkes et al. 2023)

3. (Hopefully) spur novel research on the urban development of the US.
   - Link any geolocated covariate to urban HH.
   - Study neighbourhoods in longitudinal data.

Roadmap

Motivation

Different Approaches

Methodology

Results

Validation

Summary

Different Approaches

The Ideal Geocoding Approach

The ideal approach should have high...

- Accuracy.
- Coverage.
- Relevance.

But we also want to make sure it is...

- Replicable.
- Cheap.
- Scalable.

Approach 1: Geocode Individual Addresses

Example: Aaronson, Hartley, and Mazumder (2021) geocode households from the 1910-1940 to assign them to HOLC redlining maps in 149 cities.

- **Accuracy:** Exact location for located households.
- **Coverage:** 50% to 80% of HHs.
- **Relevance:** High — Point coordinate of HH.
- **Replicability:** Low — requires access to the restricted IPUMS data.
- **Cost:** Expensive state-of-the-art geocoding API (Google Maps).
  - Back-of-the-Envelope for 1880-1950 LA: approx 3,500-4,000 USD.
- **Scalable:** Expensive but scalable.

Approach 2: Reconstruct Area Boundaries

Example: Urban Transitions Project – Logan et al. (2011) and Shertzer, Walsh, and Logan (2016) reconstruct detailed ED shape files.

- **Accuracy:** Approximate for the HH, high for the ED boundaries.
- **Coverage:** In completed cities 100%, but few cities.
- **Relevance:** Boundaries are redrawn and ad hoc.
- **Replicability:** Valuable public data resource.
- **Cost:** Expensive – immense labour effort.
- **Scalable:** Not easily scalable – not all maps survive, spread across archives, requires large team.
- Works well for specific research settings: i.e. one city (e.g. Schwank 2026) but difficult to scale.

Our Approach: Intersection Geocoding

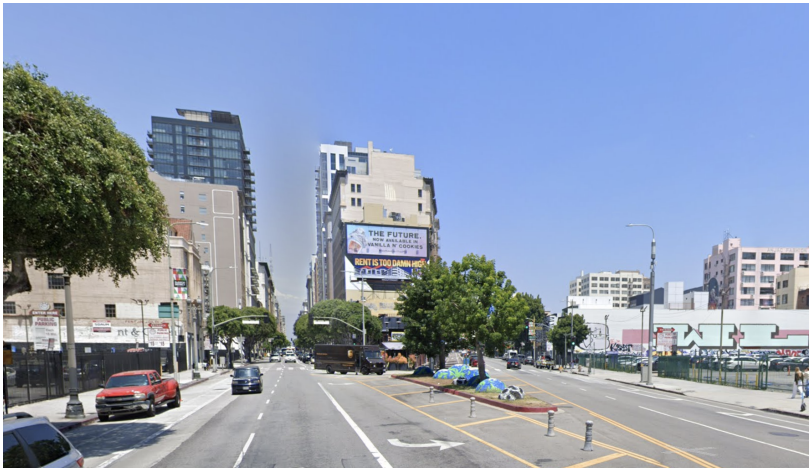
- Identify location of spatial object (ED) by locating contained features (intersections).
  - i.e. it is much easier to geocode intersections than to reconstruct the geometry of the district.

- We trade some **accuracy** for **coverage**, **replicability**, **cost**, and **scalability**.

- What would the extra accuracy buy you?
  - For most applications quite little.

Main Street & Spring Street at **(34.04, -118.23)**

in 2025



Main Street & Spring Street at **(34.04, -118.23)**

in 1880 = ED 23

What can South Main and South Spring tell us about an ED in 1880?

- Together with the location of other intersections in ED 0023, enough to geolocate it.
- Find potential intersections intersections from ED 23 in contemporary street data.

Table 1: Los Angeles, 1880: ED 0023

Intersection	Latitude	Longitude
9th street & main	34.042	-118.255
9th street & san pedro	34.037	-118.250
...	...	...
hill & jefferson	34.020	-118.275
olive & washington	34.032	-118.268
Centroid	*34.031*	*-118.262*

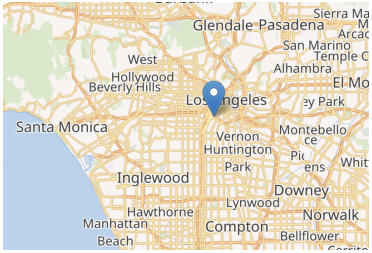


Figure 5: Approximate location of ED 0023 in 1880.

General Workflow

1. Construct all hypothetical intersections per district.
   - We do not observe real intersections.

2. Match intersections to contemporary geodata.
   - Simple string matching.

3. Collapse intersection coordinates to approximate district location.
   - Outlier detection and centroid computation.

Methodology

Validating our Approach: Simulation

Simulation Exercise:

- Monte Carlo simulations across intersection match rates.
- Measure of success is accuracy as euclidean distance between estimated and real centroid.

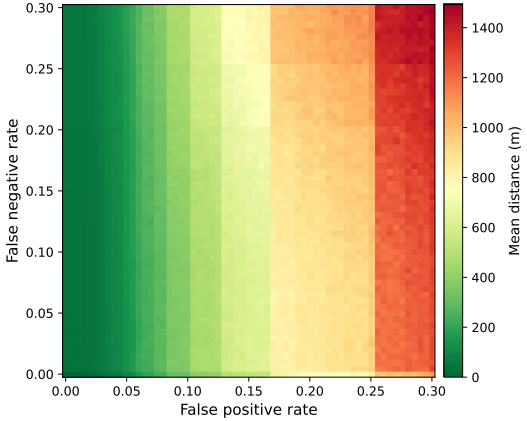
Q1: Can intersections recover ED locations?

- Are a subset of intersections sufficient to infer ED locations.
- **YES**: even 1-2 correct intersections are sufficient.

Q2: Should we prioritise maximising matches or avoiding outliers?

- Not matching to a real intersection that is in the ED vs. matching to a real intersection that is **not** in the ED.
- **False Positives**: avoid false matches.

Avoid false Matches



- FP Share: Share of real intersections that are not matched.
- FN Share: Additional false matches as a share correctly matched intersections.

Data

Full–count Census = historical intersections:

- Post-1880 enumerators recorded street names (all urban, some rural),

⇒ streets by ED (by city–year).

\+ Steve Morse Unified–ED Finder:

- Transcribe ED boundary streets from National Archives,
- Tabulated historical street-name changes,

⇒ streets by ED by census wave by city & street–name changes crosswalk.

Open Street Maps = contemporary intersections:

- Download all nodes (intersections), and connected edges (streets) from OSM,
- Construct all possible 2 street permutations: (i.e. Main–Spring–9th → Main–Spring, Main–9th, Spring–9th)

⇒ streets by intersection (with point coordinates) by city.

Geocoding Pipeline: Data Preparation

1.1 Construct all historical (hypothetical) intersections in the census data.
- We do not observe real intersections.
- Create every two-street permutation.

1.2 Extract contemporary intersections from OSM.
- Extract all two-street permutations per node (intersection).
- Include intersection point-coordinates.

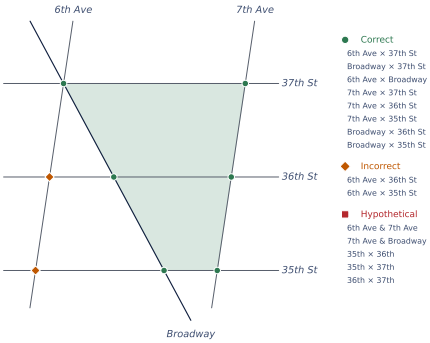


Figure 6: Example of intersection geocoding for a New York ED.

Geocoding Pipeline: Intersection Match

2.1 *Round 1:* **Exact match** on cleaned intersection key.
- Minimal string cleaning; “south main street & south spring street” ⇔ “south main street & south spring street”

2.2 *Round 2:* **Exact match** on ‘extra cleaned’ intersection key.
- Remove street types and directions; “main & spring” ⇔ “main & spring”

2.3 *Round 3:* **Fuzzy match** on ‘extra cleaned’ intersection key.
- Use a token sort ratio with a threshold of 95; “main & spring” ⇔ “man & sprig”

- Match rates are low in both directions (10-20%):
  - Hypothetical > real intersections.
  - Contemporary intersection did not exist yet & old intersections no longer do.
- Just need **sufficient** intersections to estimate ED centroid.
  - Our approach more robust to street name and layout changes.

Geocoding Pipeline: ED centroids

3.1 Spatial Clustering by ED to remove false positives (RANSAC)

3.2 Compute centroid across the remaining intersections
- Require a minimum of 2 intersections.
- Simple centroid & population weighted centroid.

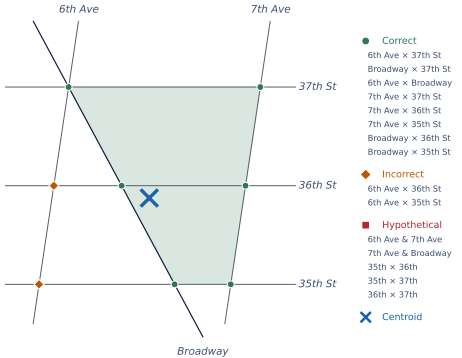


Figure 7: Example of intersection geocoding for a New York ED.

A Caveat

We do best in **median (high–)density** areas.

- *Very **low** density:* EDs are large and the centroid looses informational capacity.
- *Very **high** density:* EDs are too small (i.e. 1 - 2 streets)
  - To deal with this we dynamically merge small EDs with too few streets.

Results

Who do we include?

- All 1950 metropolitan statistical areas (MSAs) (N=185);
  - + synthetic MSAs: any county with an incorporated place of >25,000 (N=127);
  - + all rural EDs with street information.
- Total = 3,100 counties = 670 urban and 2,400 rural
- 1880: **12,352,257** (+3,062,396)
- 1950: **91,705,903** (+48,040,556)

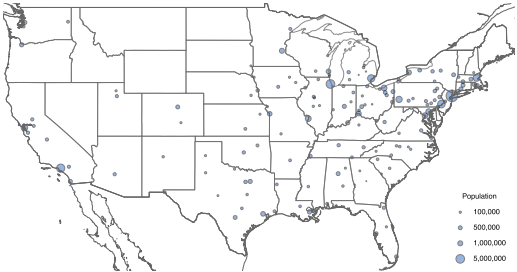
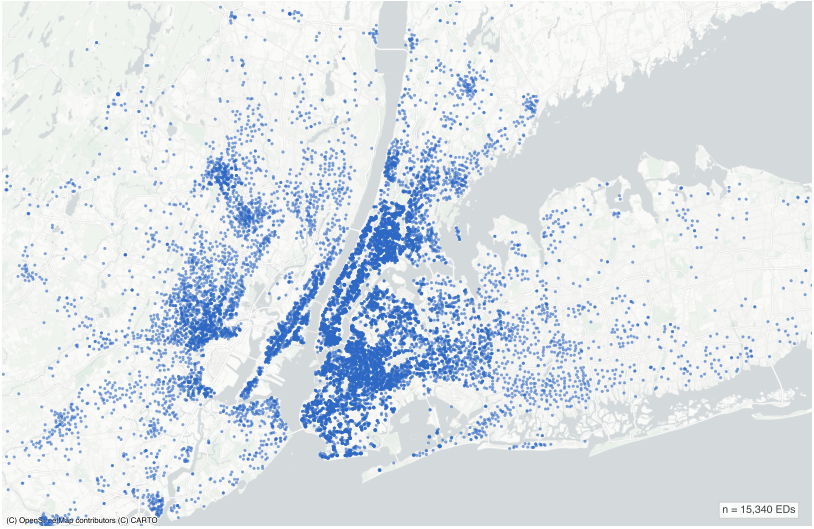
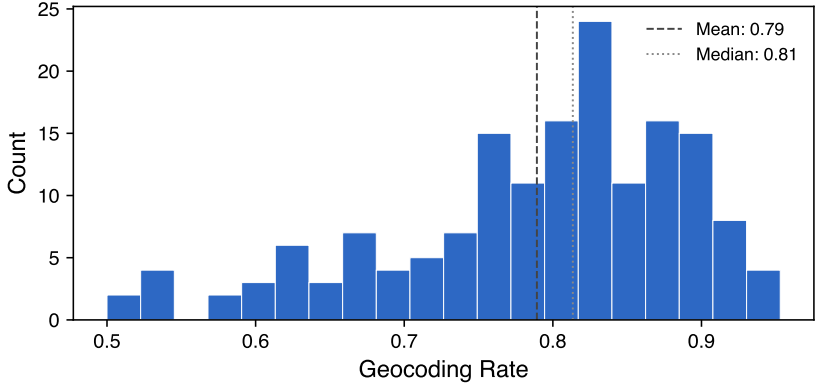


Figure 8: All MSAs and Synth-MSA in the geocoding Sample.

New York–Northeastern NJ MSA: 1950



Geocoding Rates



Where our Method falls short

THEIR WORK IS DONE

One of the house-numbering crews which placed new digits on all houses and business establishments during the last 23 months. The work is now completed.

Figure 9: The Oregonian, July 16, 1933

The Great Renaming & Renumbering of Portland, OR

Imputation

If we value **coverage** over **accuracy**, we can impute missing EDs:

Round 1: Missing ED is located between two adjacent EDs;

Round 2: Assign missing ED to location of neighbouring ED.

- Researchers using the data resource choose the trade-off.
- **>90%** of missing enumeration districts assigned.

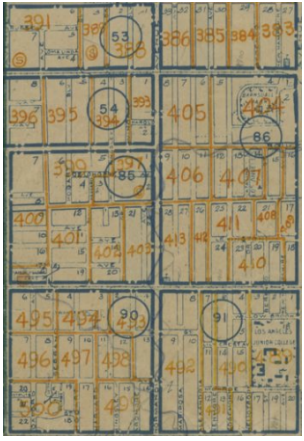
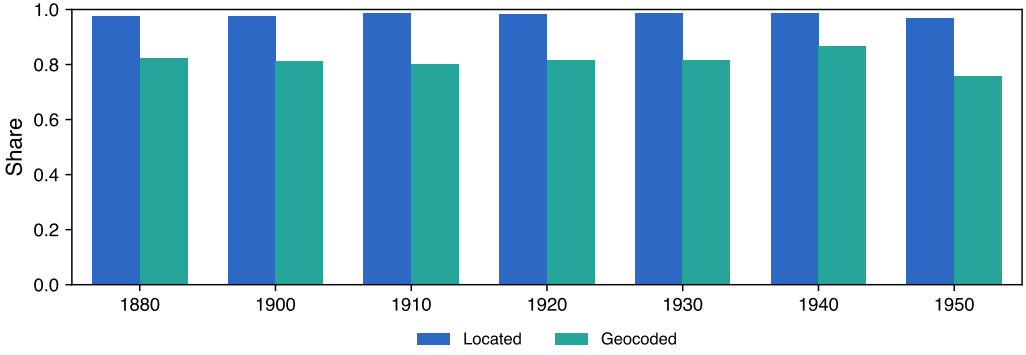
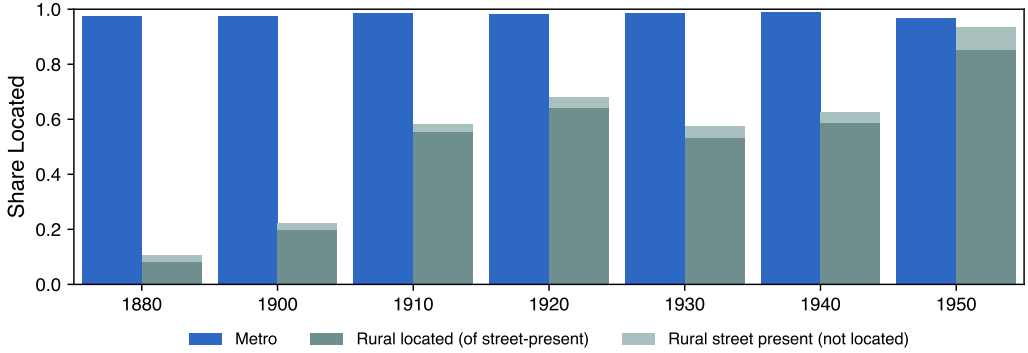


Figure 10: Excerpt from the 1940 Enumeration District Map for Los Angeles County. National Archives.

Imputation by Year

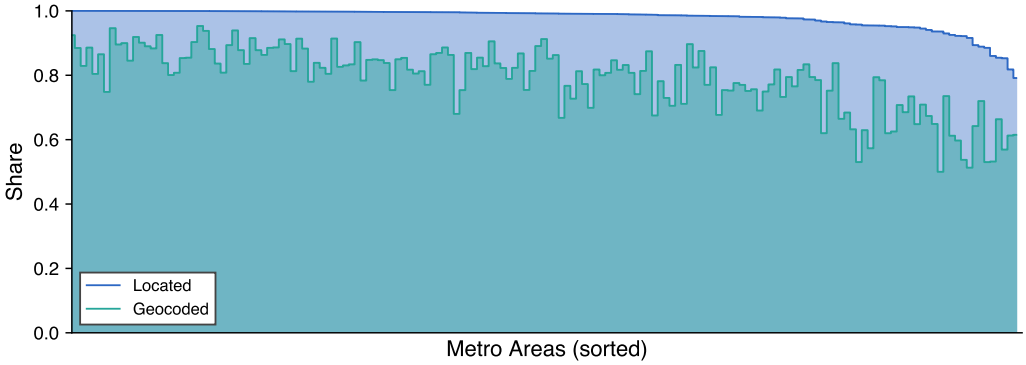


Geocoding Rates w/ Imputation

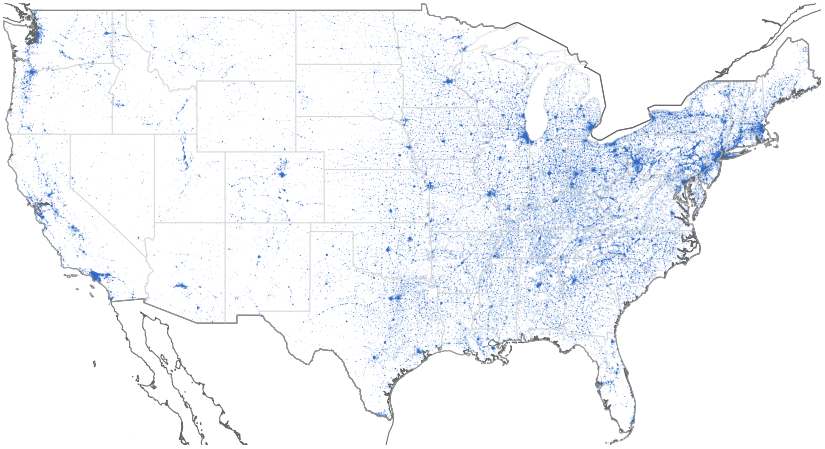


- 8,000 located in 1880 ⇒ 150,000 located in 1950.
- Metro: Located (**99.6%**) = Geocoded (**82.2%**) + Imputed (**17.4%**)

Coverage across MSAs



Enumeration Districts in Space



Internal Consistency

- We can leverage ED creation process to check the internal validity of the geocoding.
- EDs (orange numbers) were drawn onto city maps by census administrators in a sequential manner.
- ED centroid should be close to neighbouring EDs.
- Valid = within 6 nn

= **91.3%** of geolocated enumeration districts

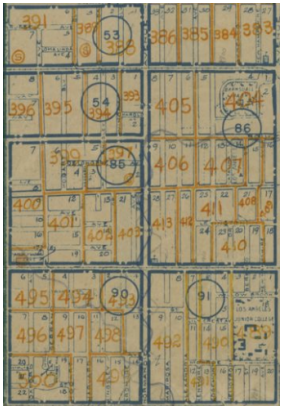
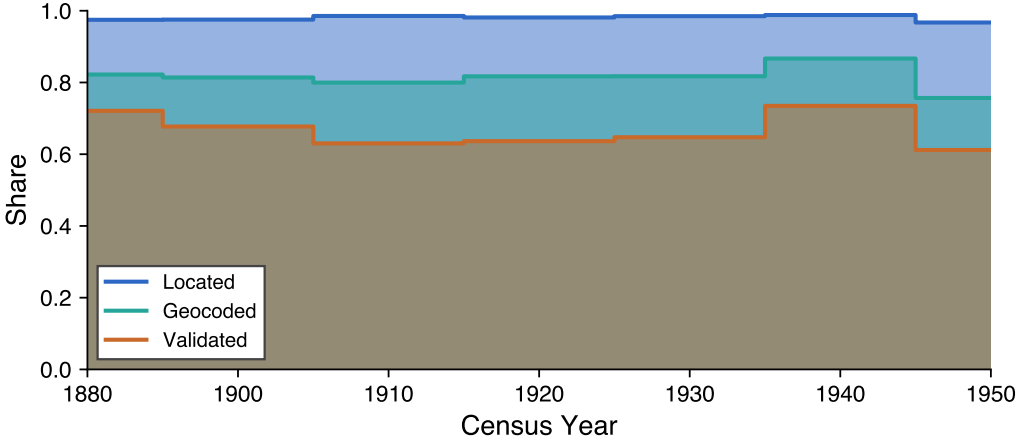
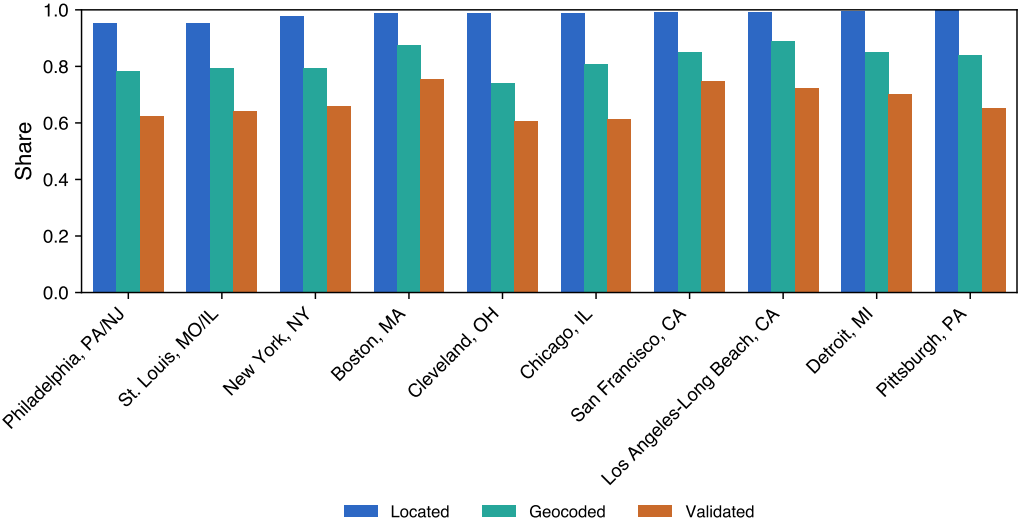


Figure 11: Excerpt from the 1940 Enumeration District Map for Los Angeles County. National Archives.

Validation over Time



Validation across Space

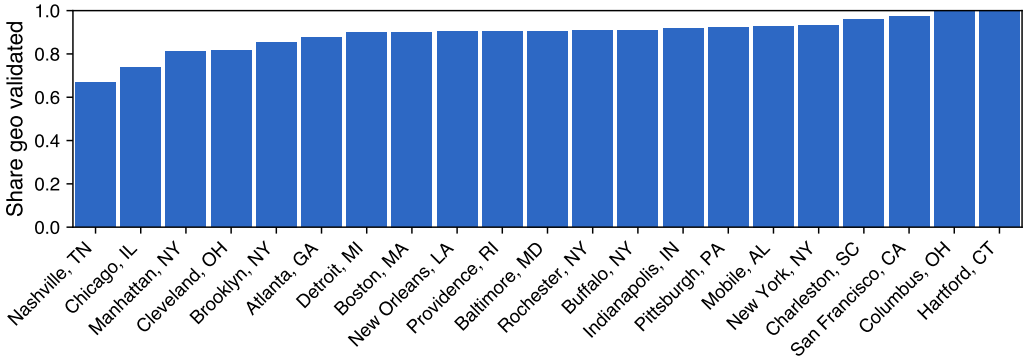


Validation

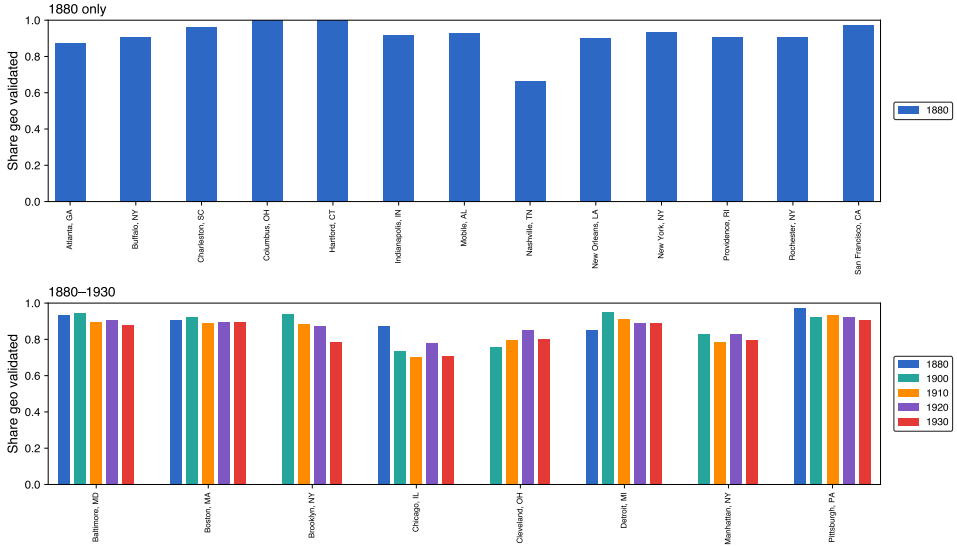
Validation: Urban Transition Project

- Urban Transition Project (UTP) has enumeration shapefiles for some years and cities:
  - 39 cities in 1880; (Logan et al. 2011)
  - 10 cities in 1900–1930. (Shertzer, Walsh, and Logan 2016)
- To validate we merge our centroid c_i to UTP polygons P_i^{UTP} and centroids c_i^{UTP}.
- ED location is valid if...
  - $d\left(c_i, c_i^{UTP}\right) \leq 250$ **or**
  - $c_i \in P_i^{UTP}$.
- **89.2%** validated,
- **76.6%** overlap,
- mean distance = **651** m.

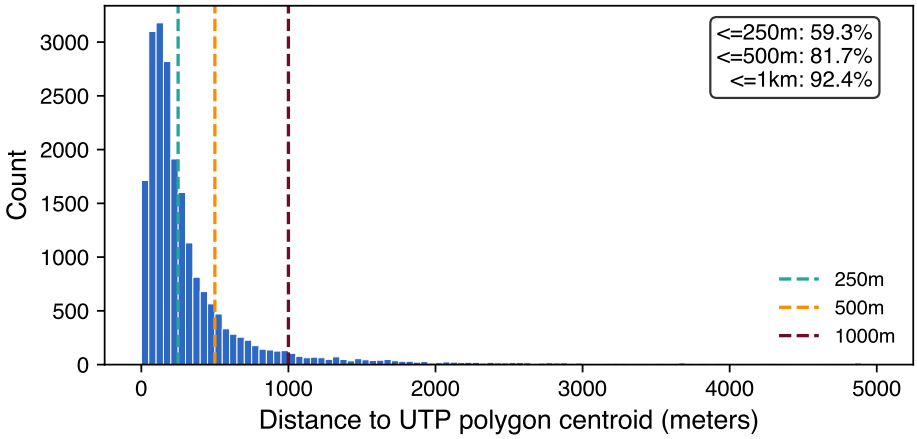
Urban Transitions Project: Pooled



Urban Transitions Project: By Year



Urban Transitions Project: By Distance



Methodology

Motivation

Different Approaches

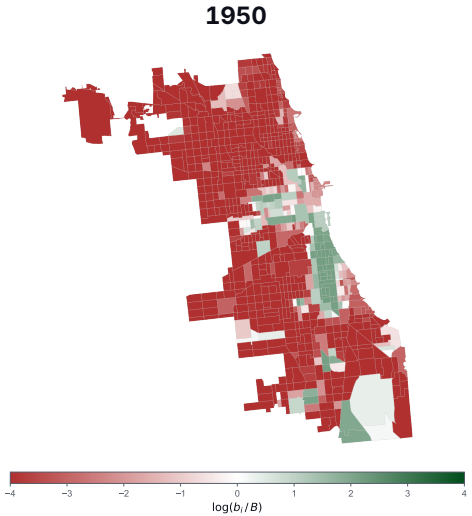
Methodology

Results

Validation

Summary

Data in Action: Racial Composition of Chicago



Conclusion

Question: Who lived where (in cities)?

- Accuracy: approximate within-city locations.
- Coverage: 97.8% of EDs in all MSAs and more.
- Relevance: fine-grained locations across time and space.
- Replicable: public data resource, transferability.
- Cheap: No expensive APIs, or large research teams.
- Scalable: Matching algorithm for >1000 city-years = approx 15 min.

Thank You!

j.p.ohler@lse.ac.uk | johannohler.com

Simulation Setup

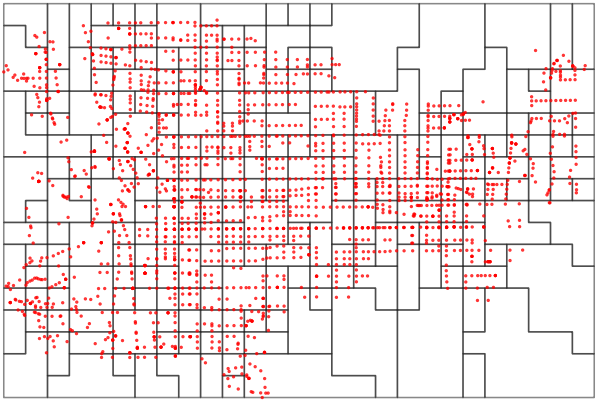
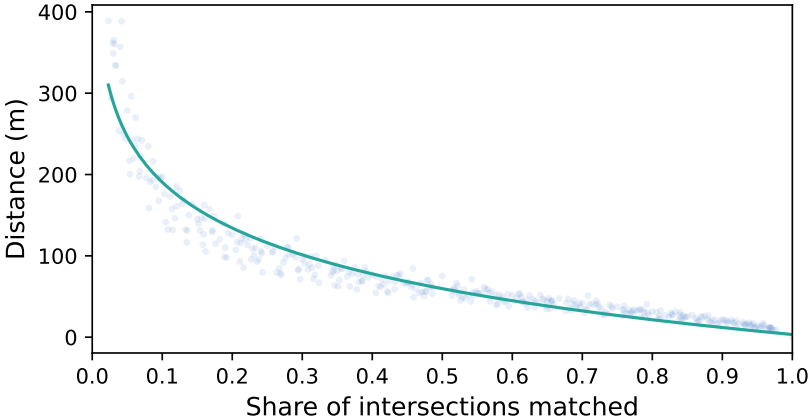


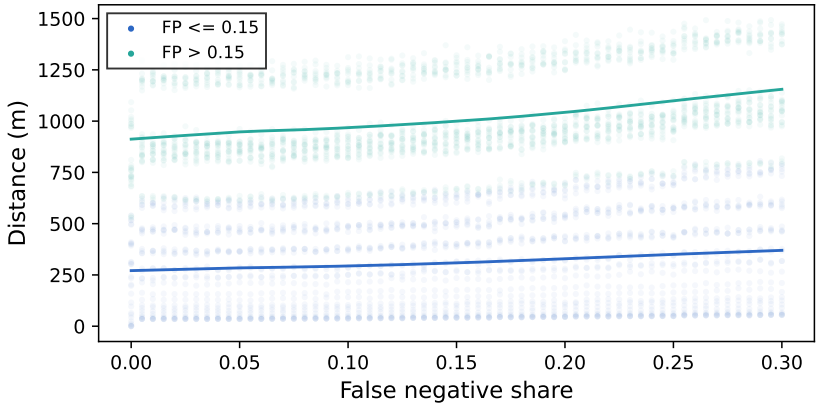
Figure 12: Real OSM intersections for Pasadena, superimposed pseudo Enumeration Districts based on 500 m2 squares.

Can intersections recover ED locations?



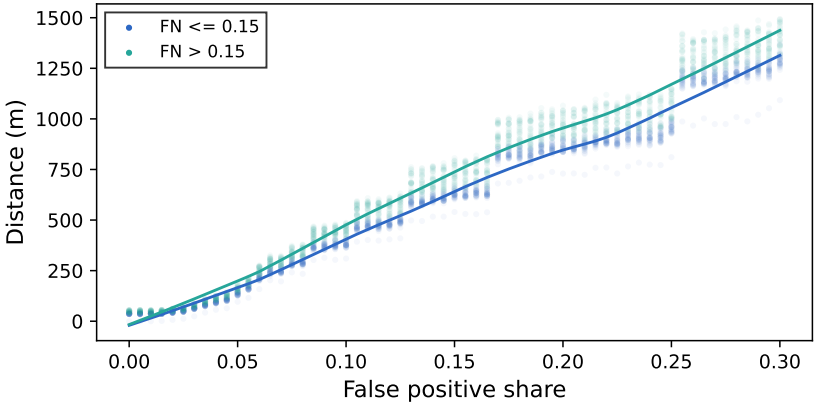
A single correct intersection is sufficient.

Missing Intersections? (*Type II* Errors)



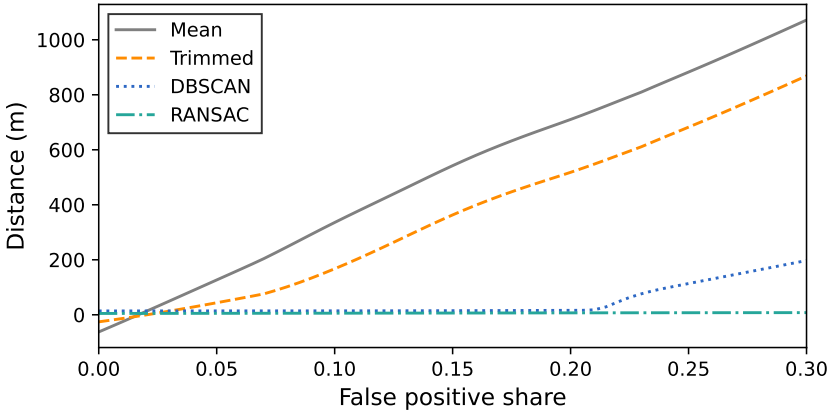
- *False Negatives*: Not matching a real intersection that is in the ED.
- *FP Share*: Share of real intersections that are not matched.

Wrong Intersections? (*Type I* Errors)



- *False Positive*: Match a *random* real intersection that is outside the ED.
- *FN Share*: Additional false matches as a share correctly matched intersections.

Outlier Detection and Removal



- *Random Sample Consensus* (RanSaC) algorithm preforms the best for outlier detection and removal.