Explorations in Economic History 87 (2023) 101477

Contents lists available at ScienceDirect

Explorations in Economic History

journal homepage: www.elsevier.com/locate/eeh

Research Paper

The census place project: A method for geolocating unstructured place names☆

Enrico Berkes[a], Ezra Karger[b], Peter Nencka[c,*]

[a] The Ohio State University
[b] Federal Reserve Bank of Chicago
[c] Miami University

ARTICLE INFO

Keywords:
Historical data
Economic history
Geocoding
Geolocation

ABSTRACT

Researchers use microdata to study the economic development of the United States and the causal effects of historical policies. Much of this research focuses on county- and state-level patterns and policies because comprehensive sub-county data is not consistently available. We describe a new method that geocodes and standardizes the towns and cities of residence for individuals and households in decennial census microdata from 1790–1940. We release public crosswalks linking individuals and households to consistently-defined place names, longitude-latitude pairs, counties, and states. Our method dramatically increases the number of individuals and households assigned to a sub-county location relative to standard publicly available data: we geocode an average of 83% of the individuals and households in 1790–1940 census microdata, compared to 23% in widely-used crosswalks. In years with individual-level microdata (1850–1940), our average match rate is 94% relative to 33% in widely-used crosswalks. To illustrate the value of our crosswalks, we measure place-level population growth across the United States between 1870 and 1940 at a sub-county level, confirming predictions of Zipf's Law and Gibrat's Law for large cities but rejecting similar predictions for small towns. We describe how our approach can be used to accurately geocode other historical datasets.

1. Introduction

The public release of full-count decennial census microdata from 1790 to 1940 has increased the quantity and quality of research studying trends and policies in the United States during this period. Much of this work uses state- or county-level data (e.g., Aaronson and Mazumder, 2011; Desmet and Rappaport, 2017; Donaldson and Hornbeck, 2016) or focuses on a small number of counties or large cities where researchers have detailed sub-county microdata (e.g., Aaronson et al., 2021; Brooks and Lutz, 2019; Fishback et al., 2020; Michaels et al., 2012; Shertzer et al., 2016). One reason for this geographic focus is data availability: commonly-used historical datasets only consistently identify states, counties, and large cities. However, states and counties cover broad geographic areas and contain important heterogeneity in demographics, policies, and access to local amenities; while larger cities can be systematically

☆ We thank IPUMS, Ancestry.com, and FamilySearch for providing us with access to data for use in this paper. We also thank participants at the Methodological Advances in the Extraction and Analysis of Historical Data conference and the Southern Economic Association conference for their helpful feedback. Any views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System. Our crosswalks can be downloaded at https://www.openicpsr.org/openicpsr/project/179401/. The code that generates these crosswalks can be viewed at https://github.com/ezrakarger/census_place_project.

* Corresponding author.
E-mail address: nenckap@miamioh.edu (P. Nencka).

https://doi.org/10.1016/j.eeh.2022.101477
Received 31 March 2022; Received in revised form 14 September 2022; Accepted 15 September 2022
Available online 20 September 2022
0014-4983/

different from their smaller peers. In this paper, we use a new method to construct consistent measures of place for U.S. residents in the 1790–1940 decennial censuses. We create a crosswalk that allows researchers using public or restricted decennial census microdata to link respondents to consistently defined sub-county locations.[1]

To construct these links, we clean and analyze raw place names from census manuscripts corresponding to cities and townships. After we identify all unique places within a given decade of census microdata, we iterate through the strings to identify a standardized location name and latitude-longitude pair. To do this, we match strings to geocoded places in NHGIS historical place point files, GNIS place files, and Google Maps.[2] When location strings are not reported or found, we exploit the existence of nearby enumeration districts to impute an accurate sub-county location. After we perform these steps for every census year in our sample, we standardize place names and locations across years to make our locations temporally consistent. We release both our code (Berkes et al., 2022b) and our resulting crosswalks (Berkes et al., 2022a) for public use.

Relative to publicly available datasets, we map many more people to sub-county locations. We match an average of 83% of the individuals and households in 1790–1940 census microdata to the longitudes and latitudes of their cities and towns of residence, compared to 23% in currently widely-used crosswalks. In years with individual-level microdata (1850–1940), our average match rate is 94% relative to 33% in widely-used crosswalks. In 1870 (the year where we obtain our best match rate), we geocode 99.1% of individuals relative to 19% in public crosswalks.

We highlight the value of our new place data with two applications. First, we take the 69,393 unique geocoded places from our census microdata crosswalks covering 1790–1940, and we iteratively cluster these places into 42,133 consistently-defined places over time. Our clustering approach is informed by the closeness and size of neighboring places and addresses the fuzziness of fixed place definitions both over time and across borders. Our clusters account for shifting place borders, annexations, subsumed suburbs, and ghost towns. We include these clusters as variables in our crosswalks, providing researchers with a data-driven definition of local metropolitan areas.

Next, we link our clusters to census microdata and create granular measures of place-level population growth over time. We show that our clustered places consistently follow Zipf's Law and Gibrat's Law for large cities in historical time periods, matching predictions from theory and modern-day empirical contexts (e.g., Gabaix, 1999; Giesen and Südekum, 2011; Ioannides and Overman, 2003). We also find sharp deviations from Zipf's Law and Gibrat's Law for small places. To the best of our knowledge, we are the first to document these patterns for smaller places across the entire U.S. in a historical context, since prior historical work focuses only on large cities, county-level patterns (e.g. Desmet and Rappaport, 2017), or a small subset of states with high-quality sub-county data (Michaels et al., 2012).[3]

Our work builds on past efforts to digitize, standardize, and improve the usability of the complete count census files. In particular, IPUMS staff have standardized large portions of the historical census enumeration sheets and created standardized variables for many common fields that researchers commonly use (Ruggles et al., 2021). Public IPUMS data generally identifies only larger cities, though the effective city-size thresholds vary over time.[4] Our approach is informed by efforts like the Census Linking Project, which provides public census data users a crosswalk that allows them to link census respondents across time, a process that otherwise would require access to restricted census data (Abramitzky et al., 2020).

Many papers (including our own prior work) use raw census strings to define sub-county areas that are smaller than IPUMS-provided cities for subsets of the historical censuses.[5] For example, Karger (2021) and Berkes and Nencka (2021) develop string cleaning methods to identify small cities and towns that had Carnegie libraries in the early 1900s. Michaels et al. (2012) standardize sub-county areas in the 1880 census and link these areas to 2000 data to study long-run trends in population dynamics.[6] Nagy (2020) standardizes cities in the 1790 to 1860 censuses to study city formation and the effects of transportation infrastructure. Feigenbaum and Gross (2021) clean city names with more than 2000 people from 1910 to 1940 to track information on telephone operators. Otterstrom et al. (2021) use linked census records from 1900–1940 to measure changes in city population size for the 1000 largest cities in 1900. Connolly (2021) digitizes locations in the 1920 census to study the impact of two-year colleges on children's adult outcomes. To our knowledge, we are the first to standardize raw location strings for the universe of all available full-count census data, allowing us to use information *across* census years to improve the accuracy of matches. Moreover, we identify additional sub-county locations for observations with missing or uninformative strings by relying on nearby, sequentially numbered enumeration districts. Finally, by publicly releasing our crosswalks and associated time-consistent clusters, we give all researchers the ability to study sub-county trends and policies and reduce duplicated effort in the research community.

[1] We define a sub-county location as any location with a finer geography than county borders, not official census-defined sub-county areas.

[2] For more details about NHGIS place point files, see Manson et al. (2021). For more details about the GNIS place files, see United States Geological Survey (2021).

[3] For a review of prior historical work on city growth rates in a historical context, embedded in a larger discussion of historical urban economics, see Hanlon and Heblich (2021).

[4] See the IPUMS documentation for a detailed description of the IPUMS "CITY" variable, which is the primary temporally consistent source of sub-county place information in publicly available census microdata. We describe the IPUMS standardization more fully and compare our mapping to theirs in Section 3.1. For 1940, IPUMS recently released a more comprehensive city variable that captures more (but not all) of the locations that we geocode. This variable is named "PLACENHG" in the public 1940 data and is constructed using NHGIS place files.

[5] It would be difficult to highlight every paper that used sub-county historical variation. In this section, we highlight a number of illustrative examples.

[6] These areas are also used in Hodgson (2018) to study the effects of the railroad on population growth.

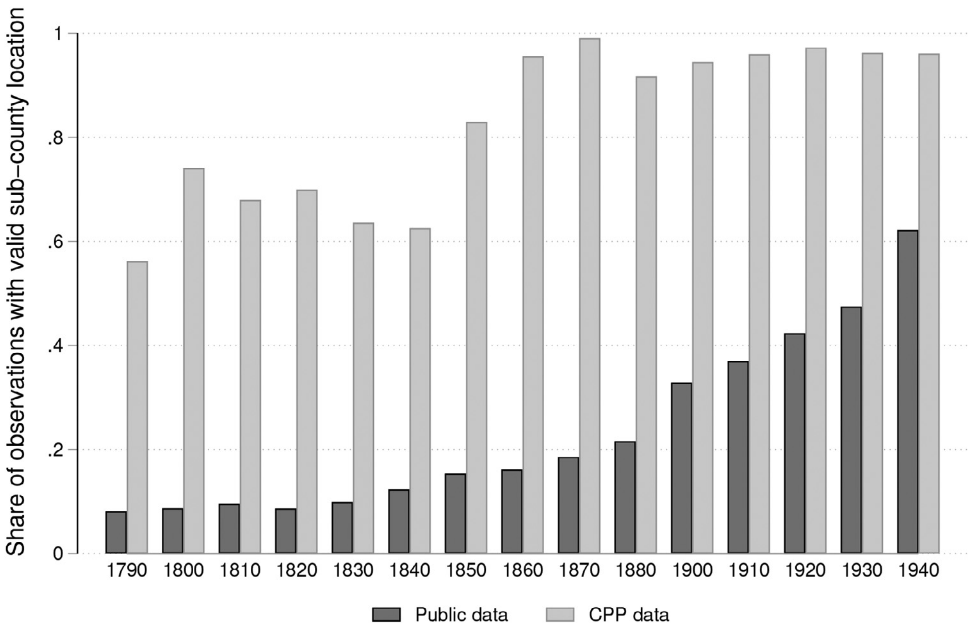

Fig. 1. Census observations with valid sub-county locations, by data source and census year. *Notes:* This figure shows the share of census observations with a valid sub-county location in each census year separately for publicly available IPUMS census data (dark bars) and the newly constructed Census Place Project data (light bars). An observation in 1790–1840 is a household. Starting in 1850, each observation is a non-slave member of the household. The 1890 census manuscripts were lost in a fire.

Relative to publicly released census data, our added value is highest when studying rural areas or smaller locations adjacent to nearby cities. By contrast, we do not attempt to geocode *within-city* locations like addresses or neighborhoods, which are the subject of recent valuable exercises focused on larger cities (e.g., Brooks and Lutz, 2019; Fishback et al., 2020; Logan et al., 2011; Shertzer et al., 2016; Aaronson et al., 2021). Sub-city geocoding requires historically accurate street layouts, which makes it difficult to apply to smaller or rural areas. Similarly, for broad rural townships, we are able to geocode respondents to a common sub-county location, but not their position *within* a township. In recent work, Ferrara et al. (2022) construct population-based crosswalks of counties and congressional districts using backward-looking population projections. Our work complements their contribution: we focus on matching individuals to geocoded places using contemporaneous location information when it is available, allowing researchers to use individual-level microdata to measure area characteristics.[7]

While our focus is historical U.S censuses, our methods apply more broadly. Along with our crosswalks, we release all code that was used to construct them. This code provides a consistent and automatic methodological approach to geocoding and clustering townships, cities, and unincorporated places. We hope that these methods will be useful in other contexts where researchers need to assign geocodes to historical documents that contain location strings. For example, U.S patents include the city, state, and county of inventors, and birth certificates often include detailed birth locations. Our methods provide an easily applicable framework for geocoding locations in these documents.

The rest of the paper is structured as follows: in Section 2 we discuss our geocoding procedure. In Section 3 we compare our geocode coverage to existing census data and show an illustrative example of how to use our data. In Section 4 we discuss how to implement our method in other applications which involve sub-county data. In Section 5 we conclude.

2. Method

In this section, we describe the method that we use to geocode the historical censuses. We discuss our approach with reference to the variables currently available in U.S. census microdata, though, as we discuss in Section 4, many of the steps below will be similar for any data involving historical locations.

We begin by identifying all the geographic information available in IPUMS' raw decennial census data from 1790–1940. These variables represent raw text strings, and, in some cases, IPUMS-standardized place names. The data includes between one and six raw

[7] We do not assign sub-county locations when we cannot reliably match a census enumeration district to a local place based on place names or nearby identified districts. This most often happens in remote areas in the pre-1860 censuses; see Fig. 1 for our match rates. Ferrara et al. (2022) discuss both the benefits and limits of their approach in Section 4 of their paper.

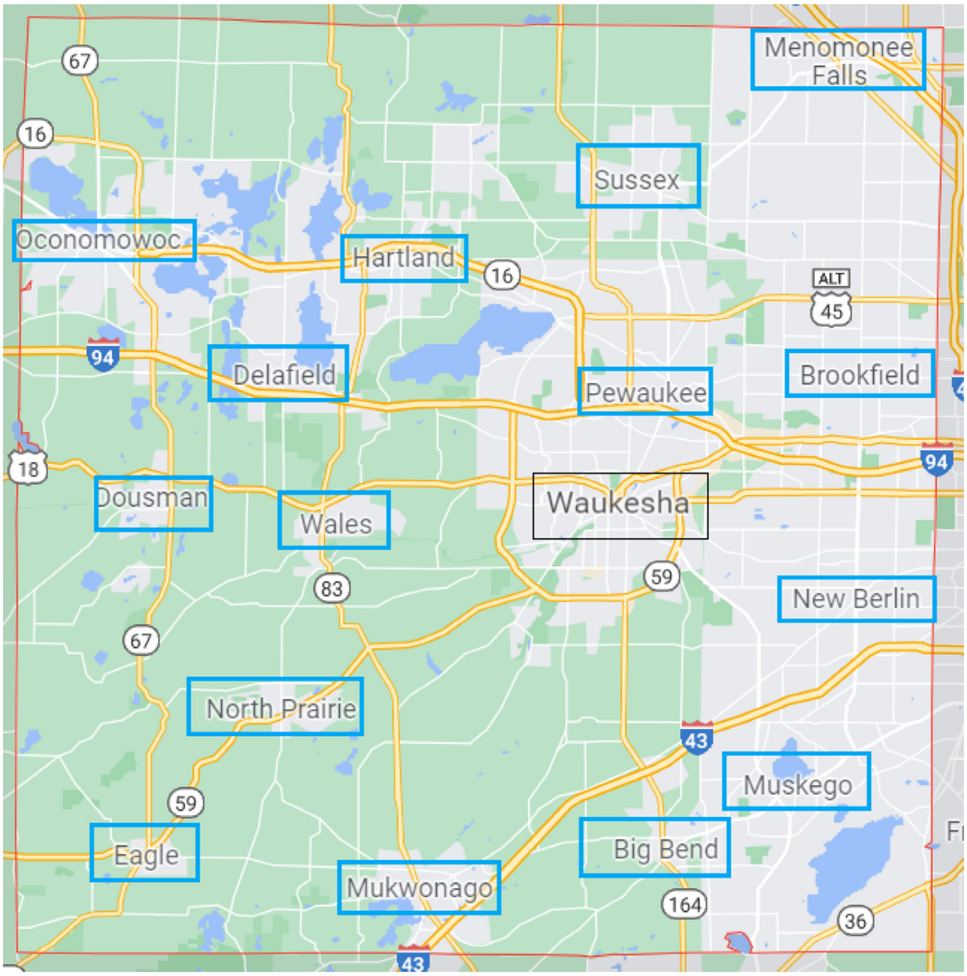

Fig. 2. Waukesha County Wisconsin, modern map with 1930 census data coverage highlighted. *Notes:* This figure is a modern-day map of Waukesha County, accessed via Google Maps. Only the city of Waukesha (highlighted in black) is identified in public 1930 census data. Our geocoded data identifies all the cities highlighted in blue, in addition to smaller areas not labeled on the map. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

location strings each year along with contemporaneous county and state identifiers.[8] These text strings are sometimes broad (e.g., "New York City"), but often contain granular information about places (e.g., "District 83, Beebe Volborg").

We clean the text strings by applying a common set of criteria, described in more detail in our Data Appendix. To summarize these steps, we standardize common prefixes, remove punctuation, remove common words (like "justice ward" or "courthouse"), and standardize cardinal directions when they refer to an explicit quadrant of a town or city. For example, the text string "Precinct 10, Aubrey [30] & Precinct 6, South Side [11]" is cleaned into the location "Aubrey."

Next, we attempt to geocode all of our clean place names in several steps, relying on historical spatial databases from IPUMS' National Historical Geographic Information System (NHGIS) and The Geographic Names Information System (GNIS). NHGIS contains the locations of incorporated and unincorporated places used by the U.S. Census Bureau from 1900 onward. GNIS is the U.S. Board on Geographic Names' consistent database of places, maintained by the federal government. We iterate over the raw census location strings, starting with the most granular place name and then using less granular place names if we cannot find a match.

For 1900–1940, we take each cleaned census place and its associated county in historical data, and we look for the most similarly named place in NHGIS that is in the correct historical county.[9] We require that the census and NHGIS strings have a match score of

8 For a full list of the variables we use, see our Data Appendix.

9 In most cases, we use county and state maps corresponding to the relevant census year. However, we have a penultimate round of matching that uses 1920 counties as our reference geography. The standardized census data uses some county or state names before they became official (e.g., West Virginia in 1860). Our baseline check would thus say that all 1860 West Virginia matches are invalid since the 1860 maps would have only had Virginia. Using the 1920 maps as a last check reduces false negative matches.

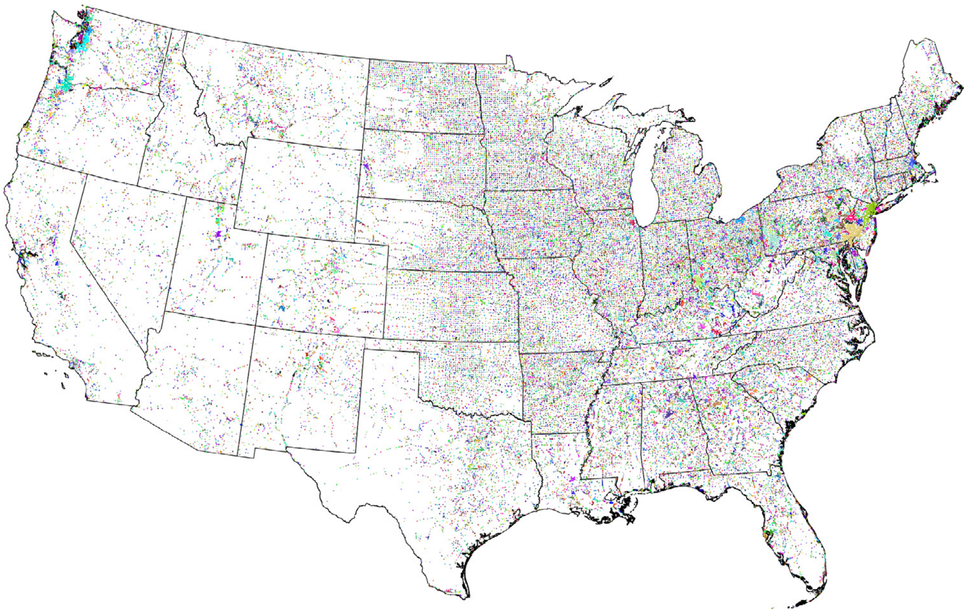

Fig. 3. All clustered places. *Notes:* This figure maps all of our 69,393 unique places across the years 1790–1940 after assigning the places to consistent clusters (with $K_{cluster} = 5$). Places within a cluster are given the same color, highlighting large colors surrounding major metropolitan areas (like New York City).

0.95 to identify accurate matches.[10] If we cannot find a match, we perform the same search within the GNIS place data, looking for matches to different feature types, ranging from populated places to post offices and valleys. For any unmatched census place after this step, we search NHGIS and then GNIS for place names in the correct county that have an edit distance of 1 with our census place. Finally, we remove cardinal directions and look again for places in the NHGIS and GNIS files with a match score of 0.95.

Our strategy for geocoding the 1790–1880 census years is similar to our process for 1900–1940. However, since NHGIS historical place points are not available before 1900, we only use the GNIS place file to match census strings to place names. The matching procedure is otherwise identical.

Once we have initial geographic coordinates (longitudes and latitudes) for places in all census years from the NHGIS and GNIS files, we complete four final steps to increase match rates and standardize those coordinates. First, we impute coordinates for places in enumeration districts when another named place within that enumeration district was successfully geocoded. Second, we impute coordinates for enumeration districts that are numerically between two successfully geocoded, nearby enumeration districts. Third, we search Google Maps for all unmatched places. We use the Google Maps latitude and longitude if it falls within the correct historical census county. Lastly, we compare across census years and standardize the spelling of matched place names and the exact coordinates of each place. This standardizes small perturbations in the reported coordinates of places across NHGIS, GNIS, and Google Maps. In our crosswalks, we include flags that indicate at which step each match is made so that researchers can exclude these imputed matches as desired. The Data Appendix includes more details on all of these steps.

After this procedure, there are 708,928 unique census year -by- cleaned location observations that we extract from the raw census data. Of those year-place observations, we fail to geocode 42,155 places (6% of the total number). 390,913 places (55%) match to NHGIS places in our first attempt, and an additional 198,491 (28%) match to the most common types of GNIS places in our first attempt. 14,690 (2%) match to NHGIS and GNIS places using slight variation in the fuzziness of match requirements, and an additional 54,403 (8%) match through our two enumeration district imputation steps. By ensuring time-consistency of identically-named place names across census years, we geocode an additional 7357 (1%) of places. Lastly, we geocode 919 (0.2%) places using Google Maps.

Our crosswalks provide consistent longitudes and latitudes for each person's town, city, or unincorporated place of residence. To increase the usability of our crosswalks, we also assign modern-day county and state identifiers to each geocoded place using 2016 United States County and State shapefiles.[11] This provides a temporally consistent measure of county and state of residence for all geocoded observations. Our crosswalks provide an accurate and consistent way to identify the county of residence for the vast majority of U.S. residents, complementing recently-constructed spatial harmonizations of changing county borders over time (Ferrara et al., 2022; Hornbeck, 2010; Perlman, 2014).

[10] The match score is calculated using the string-matching process of the fuzzywuzzy library in Python. It combines several methods for calculating a measure of 'distance' between potential strings, normalizing by string length.

[11] Historical county identifiers are available in the public census data, but county boundaries can change over time.

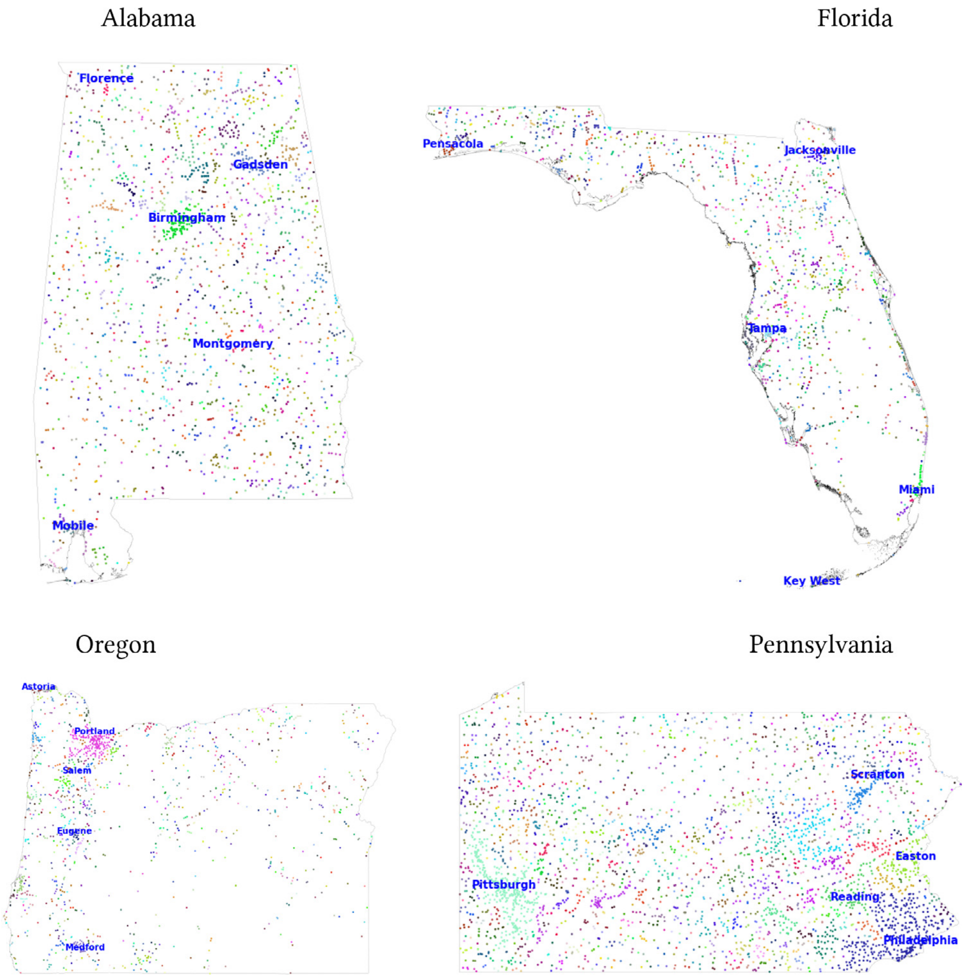

Fig. 4. State-level maps of clustered places. *Notes:* This figure maps our geocoded places in four states: Alabama, Florida, Oregon, and Pennsylvania. We highlight the five largest clusters in each state (with $K_{cluster} = 5$).

3. Results and application

3.1. Geocoding rates

In this subsection, we describe the coverage of our geocoded data across decades and compare our match rates to existing public data. In all years, our match rates allow researchers to observe more sub-county locations relative to previously available sources. To see this, we calculate the share of observations in public census data that can be geocoded using extracts from IPUMS. In particular, we compare our cities to observations with non-missing IPUMS standardized city variables “CITY” or (for 1940 only) “PLACENHG.”

We show our match rate comparison in Fig. 1, which plots the percent of census observations with valid sub-county locations across years for both our crosswalks and publicly available IPUMS data. We successfully match 56–99 percent of observations to a sub-county location, depending on the census year. Our match rate generally increases over time, particularly when the census moves to the collection of individual-level data in 1850. Match rates surpass 91% in all years from 1860 to 1940.[12] Our match rate differs across years due to the various methods that censuses used to collect information on geographical locations. For example, the 1870 census has significantly more geographical digitized information relative to prior years.

When are we unable to make a match? In the vast majority of cases, our failure to link a respondent to a valid latitude and longitude occurs when there is *no* geographical information beyond the county recorded on census forms for a given enumeration district, and we have limited information on adjoining enumeration districts. For example, some remote districts in earlier census years have no

[12] In 1790–1840, each observation corresponds to a household. Starting in 1850, an observation is a person. We focus on census *observations* and not the count of people throughout this paper because the decennial population censuses did not collect information about the slave population in all years.

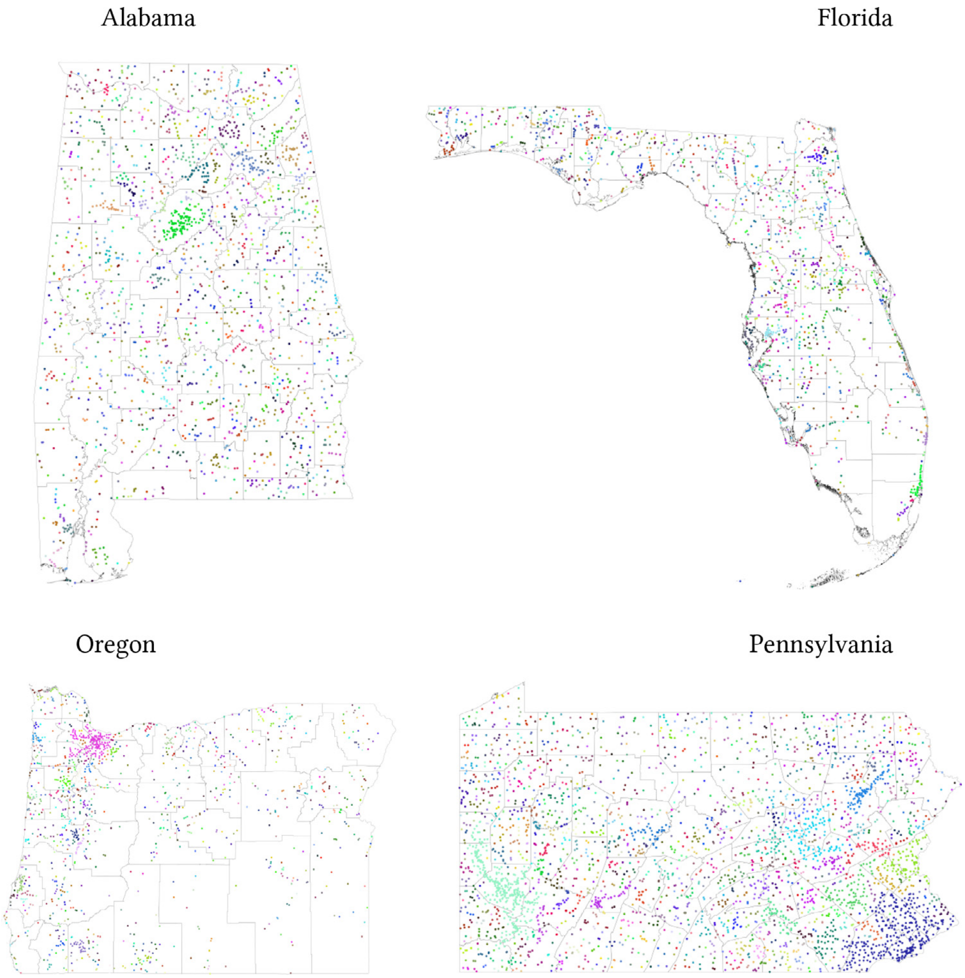

Fig. 5. State-level maps of clustered places with county borders. *Notes:* This figure maps our geocoded places in four states: Alabama, Florida, Oregon, and Pennsylvania. We highlight the county borders in each state to emphasize the granularity of our geocoded places relative to the larger counties.

informative name attached to them, simply bearing names like "District 10." Conceptually, it is possible to map these places manually by consulting the original enumeration district maps, as is done by Connolly (2021) for a set of 1920 places. Unfortunately, since enumeration district boundaries change over time, this procedure would need to be repeated for every decade in our sample. These unnamed locations tend to be small rural areas. Our crosswalks focus on places where we have a high degree of certainty that we can assign a valid sub-county location using contemporaneous location information.

3.2. Example: Waukesha county

To illustrate the value of our data and approach, consider Waukesha County in Wisconsin. Today, Waukesha County is Wisconsin's third-largest county by population and covers 581 square miles. The county is geographically diverse: its eastern portion is an extended suburb of Milwaukee and is heavily commercialized with manufacturing and service industries, while the western and southern portions are rural and contain significant farmland.

In 1930, Waukesha County had 52,000 people. In publicly available 1930 census data, the only available sub-county location is the county seat, Waukesha.[13] In 1930, the city of Waukesha contained roughly 35% of the county population, leaving 65% of the city lacking a valid sub-county location. By contrast, we assign *100%* of the 1930 Waukesha County population to a valid sub-county location.[14]

13 This location starts to be identified in the public census microdata from IPUMS in 1900. Before this, there is no city identified in this county in the public census data.

14 Our match rates for this example are similar if we focus on other census years with individual-level microdata. In 1940, IPUMS data also captures some of these sub-county locations with their newly constructed "PLACENHG" variable.

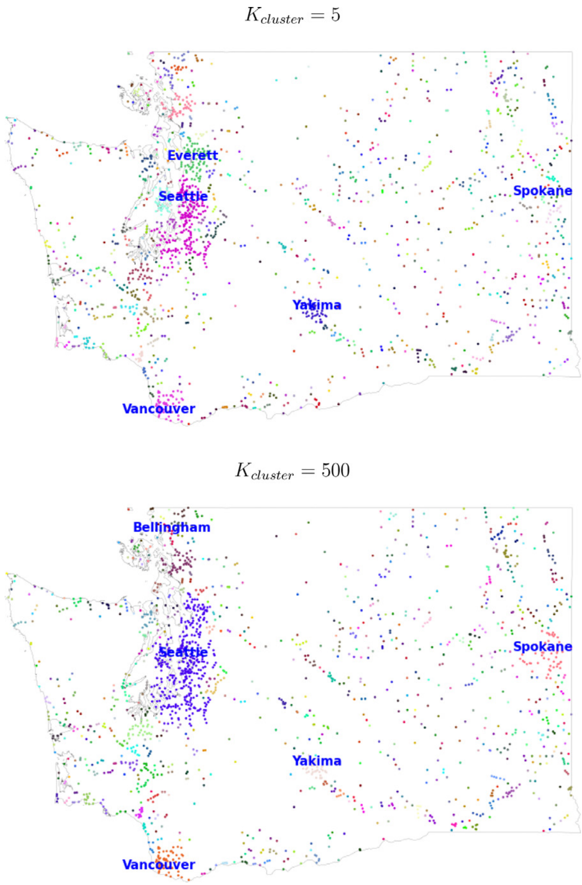

Fig. 6. Clusters in Washington State with different levels of $K_{cluster}$. *Notes:* This figure maps our geocoded places in Washington with more conservative clustering ($K_{cluster}$ = 5) and more aggressive clustering ($K_{cluster}$ = 500). The largest difference is in the implied size of the Seattle metropolitan area.

In Fig. 2, we illustrate the coverage of our geocoded 1930 places using a map of modern-day Waukesha County locations. The city of Waukesha—framed in black—is the only sub-county 1930 location in public census data. By contrast, all of the cities highlighted in blue are also included in our new geolocated place data. This includes extremely small locations in 1930—for instance, Wales had only 123 residents, while Dousman had 256 residents. Fig. 2 highlights that our new data shows the geographic diversity of Waukesha County. Even in 1930, the jobs and industries of the eastern portions of the county were changing and becoming more industrialized relative to the more rural parts of the county. With our location data linked to publicly-available IPUMS microdata, researchers can observe and study the persistence of these within-county differences throughout history.

3.3. Clustering

While our place-level longitude-latitude mappings are comprehensive, these places are not defined consistently over time. To address this limitation, we use an iterative density-clustering approach to map our 69,393 unique locations across the 150-year period of 1790–1940 to a smaller set of 42,133 consistently-defined places. While most places are distinct and not combined in this step, larger cities like Atlanta, Pittsburgh, and Chicago have borders that expand over time as they merge with other towns and experience high levels of in-migration. Our method captures these expanding borders, allowing us to consistently measure features of these cities over time.[15]

[15] We cluster places consistently over time so that researchers can use our cluster identifiers to track city growth, shrinking urban borders, mergers, and annexation. For example, in 1907 Pittsburgh annexed nearby Allegheny City against the wishes of a majority of Allegheny residents who, in repeated referenda, rejected the annexation attempt. The annexation was forced on Allegheny by Pennsylvania, whose legislators passed a law allowing a majority of the combined voters from Pittsburgh and Allegheny City to determine the results of the annexation, even if a majority of the voters in the targeted city (Allegheny City) rejected the annexation attempt. For more discussion of this and related annexations, see Lonich (1993). Researchers interested in the distinction between Pittsburgh and Allegheny City can use our data to differentiate those two places in years where

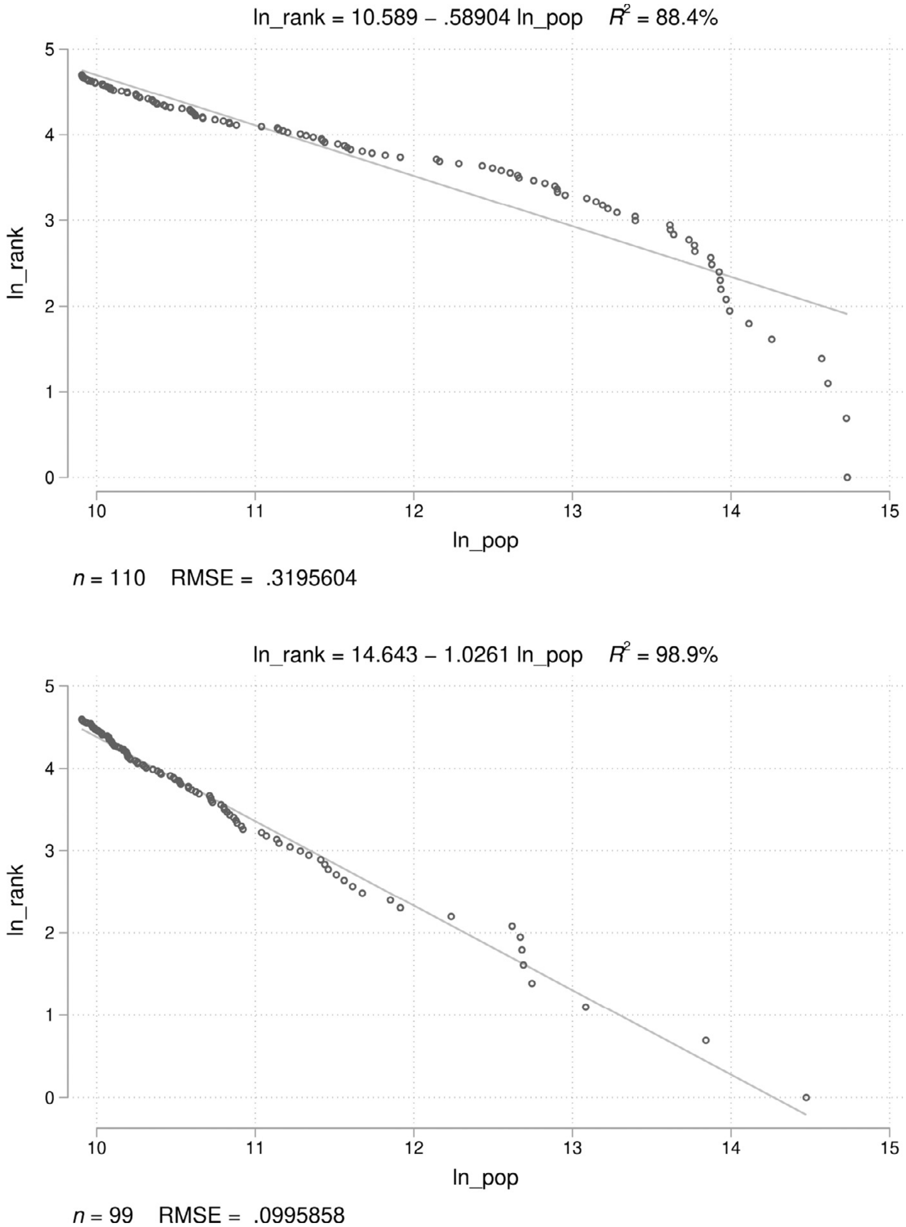

Fig. 7. Zipf's Law for 1870 places with population 20,000+ (CPP vs. IPUMS). *Notes:* This figure plots log(population) vs. log(rank) for all cities with population greater than 20,000 in 1870 labeled by IPUMS (top panel) and in our geocoded places (bottom panel).

We iteratively cluster any neighboring places that are within three miles of each other within census years and across census years. We also cluster together any two places i and j that are within $100 * K_{cluster} * max\{sharepop_i, sharepop_i\}$ miles of each other, where $sharepop_i$ is the fraction of the population in decennial census data that we map to place i across all years (1790–1940), assigning equal weight to each year. For our main results, we rely on a constant $K_{cluster}$ of 5. This choice of clustering allows large cities to be combined with more of their suburbs. For example, Chicago contains roughly 2.5% of the people in decennial census microdata from 1790 to 1940. When $K_{cluster} = 5$, Chicago will be combined with any smaller place within 12.5 miles of it. We cluster places consistently over time by defining each cluster to be the connected component of all close neighbors for each place.

In Fig. 3, we map all of our consistently defined places. We color each cluster to reflect close neighbors. Our individually-geocoded places reflect the geographic distribution of people in the United States from 1790 to 1940, highlighting the densely populated New

the places were enumerated distinctly. But as the cities merged, the definition of Pittsburgh and Allegheny became amorphous, and our clusters provide a time-consistent approach to measuring the number and type of people living in the Pittsburgh area over time.

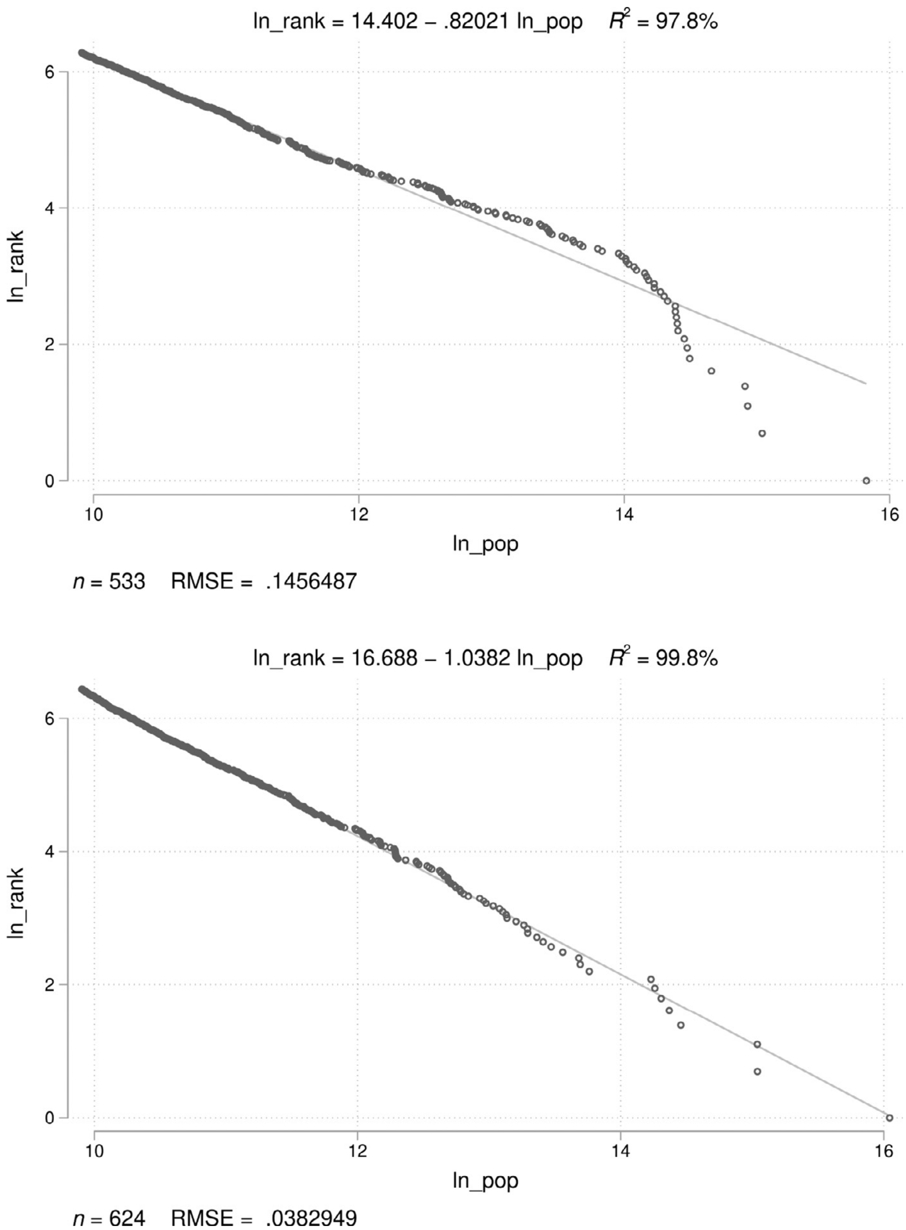

Fig. 8. Zipf's Law for 1940 places with population 20,000+ (IPUMS vs. CPP). *Notes:* This figure plots log(population) vs. log(rank) for all cities with population greater than 20,000 in 1940 labeled by IPUMS (top panel) and in our geocoded places (bottom panel).

England corridor. Our clusters show consistently-defined metropolitan areas that can spread across state and county borders. We include these cluster definitions in our crosswalks so that other researchers can use them.

To further show the value of our clusters, in Fig. 4 we focus on four representative states across different census regions and highlight the individual geocoded places, our consistently-defined clusters, and the five largest clusters in each state. The top left panel shows that our clustering combines the Birmingham, AL suburbs into one cluster. In the top right panel, Miami and Tampa Bay, FL, are combined into distinct clusters with their nearby suburbs. In the bottom panels, the dense west coast of Oregon is clustered into a small number of metropolitan areas. In Pennsylvania, Pittsburgh and Philadelphia form distinct clusters.[16]

In Fig. 5, we show the same four states with modern-day county borders. These maps show the value of our geocoded places to researchers who want to analyze spatial variation in access to policies or spatial outcomes in a historical setting. We geocode an average of 23 places per modern-day county in the U.S. This granular classification of locations gives researchers the tools to

[16] Figures A1–A4 show full-page versions of these state-level figures for easy readability.

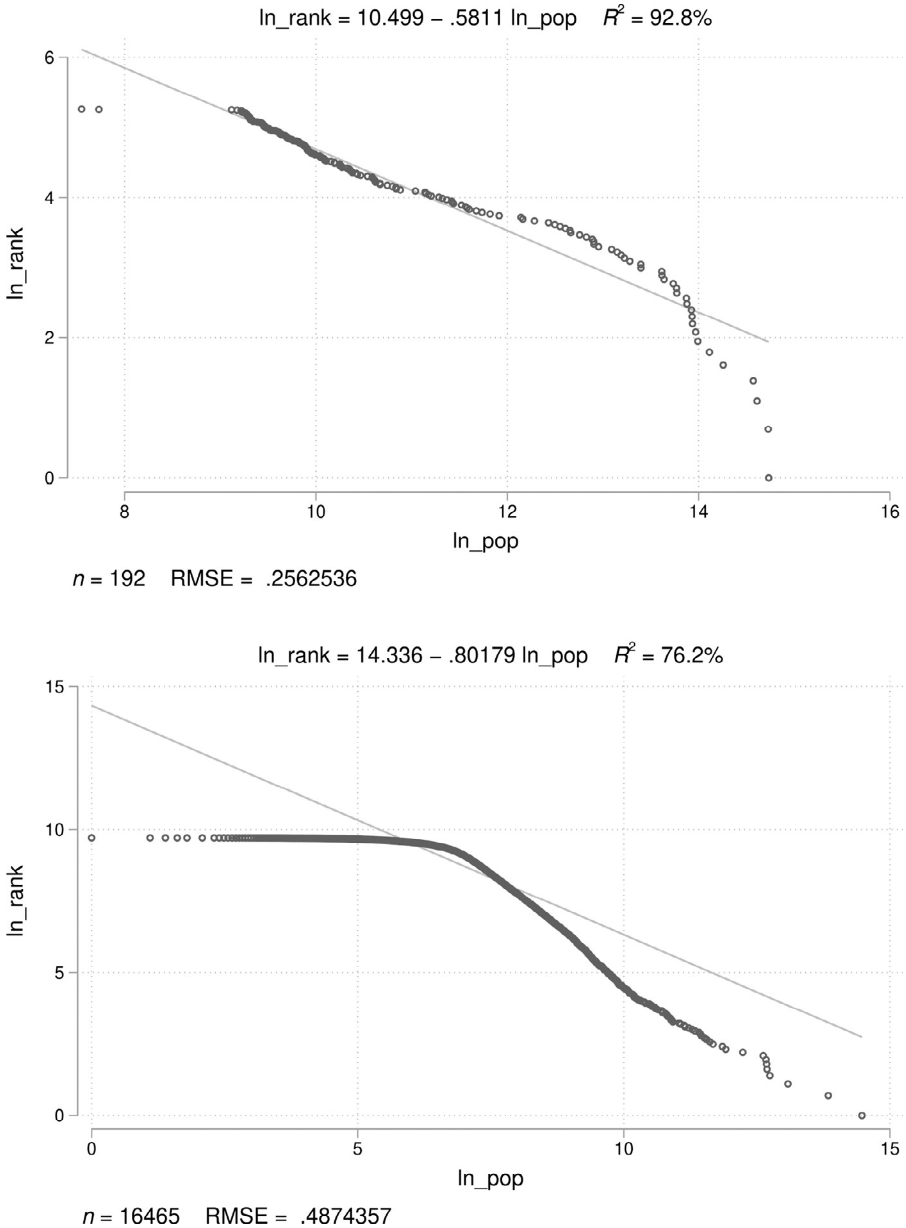

Fig. 9. Zipf's Law for all places in 1870 (IPUMS vs. CPP). *Notes:* This figure plots log(population) vs. log(rank) for all cities in 1870 labeled by IPUMS (top panel) and in our geocoded places (bottom panel).

analyze distance-based access to local programs, within-county border discontinuity designs, and urban/rural migration patterns within counties.[17]

Our preferred value of $K_{cluster}$ in the above specifications is 5 because that allows large cities to subsume close suburbs, but it maintains the geographic distinctness of nearby small places. We also provide alternative clusters for different values of $K_{cluster}$ so that researchers can choose the clustering that best fits their analysis based on the level of geographical variation in which they are interested. In Fig. 6, we show our clusters for Washington with two levels of aggregation: $K_{cluster} = 5$ and $K_{cluster} = 500$. With the less aggressive level of aggregation, Seattle and Everett (29 miles away from each other) form distinct clusters with their respective surrounding suburbs and nearby towns. With more aggressive clustering ($K_{cluster} = 500$), the Seattle and Everett clusters merge and become one larger Seattle metropolitan area, and the Spokane cluster absorbs more nearby towns.

[17] Figures A5–A8 show full-page versions of these state-level figures for easy readability.

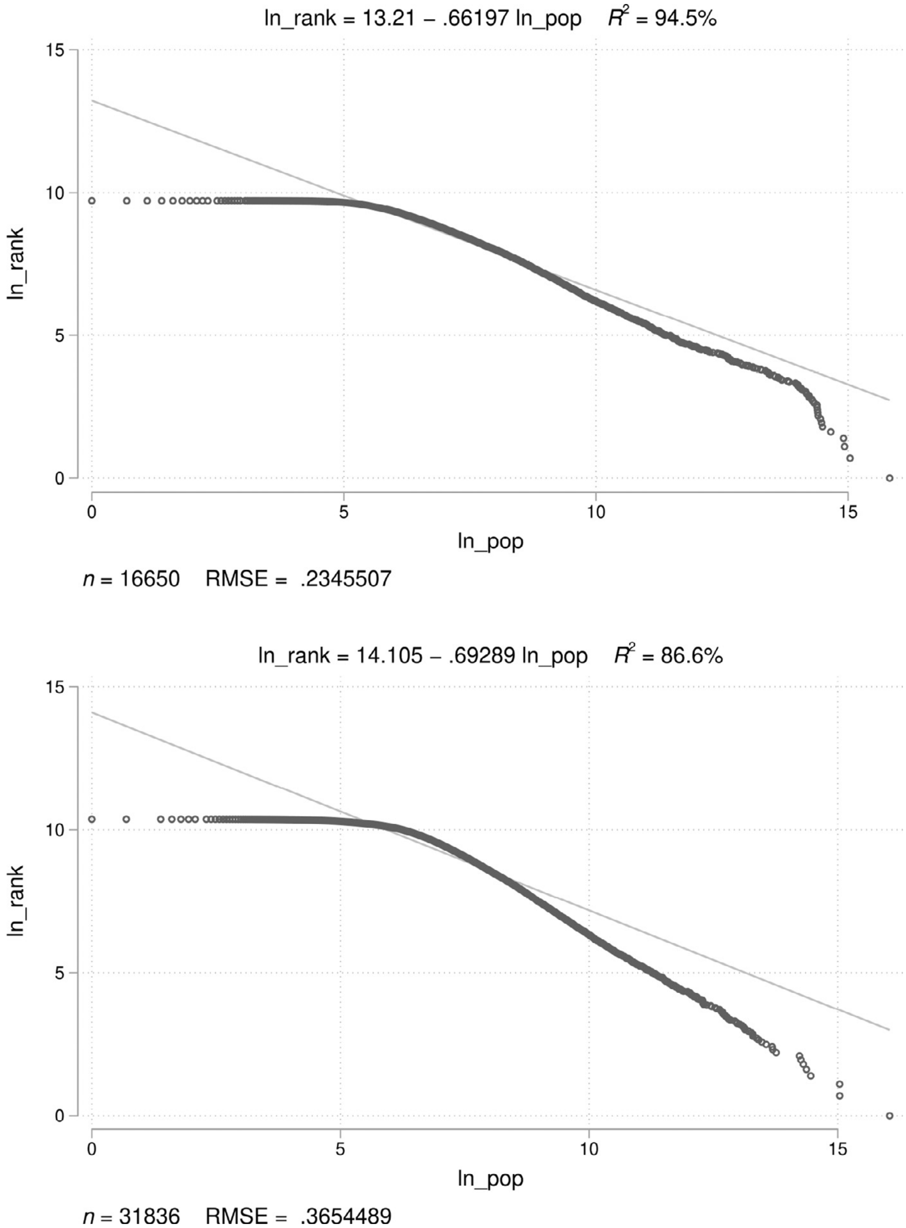

Fig. 10. Zipf's Law for all places in 1940 (IPUMS vs. CPP). *Notes:* This figure plots log(population) vs. log(rank) for all cities in 1940 labeled by IPUMS (top panel) and in our geocoded places (bottom panel).

3.4. Population dynamics

To showcase our new crosswalks, we use our preferred clustered places to examine the size distribution of historical places in the U.S. and place-specific population growth. A large literature in urban economics models and empirically quantifies the formation and growth of cities. This literature focuses on two "laws": Zipf's Law—which states that there is a linear relationship at the city-level between log(population) and log(rank) of cities in terms of population—and Gibrat's Law—which states that the population growth rate is independent of city size. Gabaix (1999) showed that if a set of cities grow independently of their initial city size, following Gibrat's Law, then the steady-state size distribution will follow Zipf's Law. Testing these predictions using high-quality historical data gives urban economists the ability to evaluate theories of long-run city population dynamics.

We use our geocoded historical places to evaluate these two laws. In Fig. 7 (for 1870 places) and 8 (for 1940 places), the top panel shows the relationship between log(population) and log(rank) for IPUMS-defined cities and the bottom panel shows the same relationship for our geocoded places. We focus on places with populations over 20,000. In the IPUMS-based top panels, there is

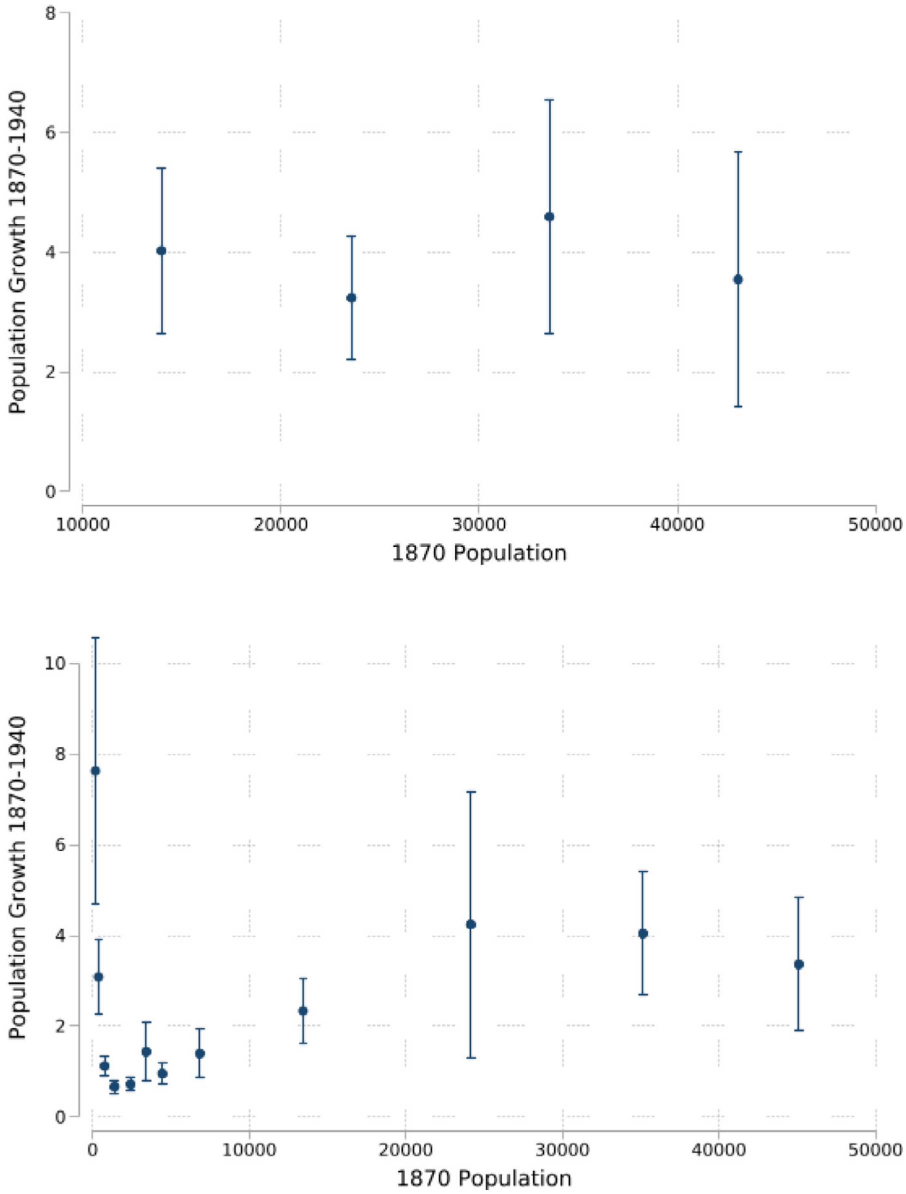

Fig. 11. Gibrat's Law (IPUMS vs. CPP). *Notes:* This figure tests Gibrat's Law by plotting the relationship between population growth rates between 1870 and 1940 (y-axis) and baseline 1870 population levels for places labeled by IPUMS (top panel, x-axis) and in our geocoded places (bottom panel, x-axis). We plot binscattered growth rates. We focus on places with population less than 50,000 to emphasize the difference between the two approaches for small cities and towns. The y-axis units is percent growth, so a value of 2 indicates that a place tripled in size from 1870 to 1940.

a deviation from Zipf's Law at the right tail of the city size distribution in both 1870 and 1940—the most populous cities look smaller than what Zipf's Law predicts if we use IPUMS's city classification. In the bottom panel of each figure, we plot the same log(population)-log(rank) graph with our geocoded and clustered places. Our decision to cluster places combines large cities with their suburbs, and these clustered places match the predictions of Zipf's Law.

In Figs. 9 and 10, we extend Zipf's Law to all tracked places in 1870 and 1940 respectively. In the top panels, we show the IPUMS-based log(population)-log(rank) plot, and in the bottom panels, we present the log(population)-log(rank) plot for all of our geocoded clusters. In both years (1870 and 1940), we see sharp deviations from Zipf's Law for places with fewer than 500 residents in our geocoded data. Focusing on the smallest places, we see that they are underrepresented in our sample of places relative to the predictions of Zipf's Law. In 1870, we do not see a similar pattern in IPUMS's data because the publicly available crosswalks do not contain the locations of small cities and towns. In 1940, when IPUMS coverage is higher, we do see the same deviation from Zipf's Law for smaller cities and towns.

In Fig. 11, we compare the prediction of Gibrat's Law in IPUMS' city data (panel A) and our geocoded clusters (panel B). We focus on place-level population growth from 1870 to 1940 for places with populations less than 50,000 in 1870.[18] IPUMS consistently tracks very few cities over this time period, but for the 144 cities with a population over 10,000 in 1870, we see a close match to the predictions of Gibrat's Law—city growth from 1870 to 1940 seems uncorrelated with the initial 1870 population. Cities of all sizes quadrupled in size between 1870 and 1940. But in our geocoded places we see a strikingly different conclusion. We find that population growth rates were U-shaped over this 70-year period. The population of smaller places grew almost nine-fold on average, while the population of the largest places more than quadrupled. But for mid-sized places with populations around 1000 people, we see much lower growth rates around 150%. This is a pattern that also exists at the county level, as shown by Desmet and Rappaport (2017). We are the first to show these historical patterns at the town- and city-level for all places in the United States.

4. Discussion and application to additional datasets

So far, we have discussed our method and applications exclusively in the context of the historical U.S. decennial censuses. However, much of the code that we release can be used to geocode other sources of historical data. In addition, the steps that we use can serve as a conceptual guide for other researchers. However, like all methodological contributions, our approach will apply well in some cases and less well in others. In this subsection, we discuss these tradeoffs.

Our current code was developed to geocode historical U.S. cities and townships. It is most easily applied to other documents that contain the same level of geographic detail. Two important examples are patent documents and birth/death certificates. Both typically contain the city of the invention or event, and both—because of the limits of modern OCR software and historical record-keeping practices—are often measured with some error. Our code can clean these location strings and fuzzily match them to historically accurate geocoded places. All that is required as an input is a vector of possible strings for each document. Since our code was developed using U.S. census data, it may need to be modified: For example, in the censuses, there are predictable strings that need to be cleaned before geocoding (e.g., "WARD", "DISTRICT"). To the extent that different datasets contain different strings that need cleaning, this portion of the code should be adjusted accordingly. Otherwise, our procedure has few census-specific steps.

By contrast, because of our setting, our code cannot directly geocode sub-city data (e.g., streets or blocks) or locations outside the U.S. However, with appropriate changes, the structure of our code could be used in these cases as well. For example, to geocode streets, one could take information on raw streets in the censuses, clean them using our algorithms, and match them to contemporary street databases (instead of NHGIS/GNIS place points). Similarly, our approach and code can be generalized to other countries, though many methodological choices will need to be made as a function of the underlying strings and location databases available for that country.

Beyond our code, our geocoded census locations can also be used in methodological projects. For example, researchers linking birth certificates to decennial census microdata face the limitation that within a state or county, there can be multiple people with similar names. Using our geocoded sub-county locations to block on geographical information and discard unlikely links can increase match rates and lower false positive rates.

5. Conclusion

Researchers often use place-level data to measure the causal effects of local policies and to describe historical trends. These analyses are often done at the county- or state-level because it is challenging to link individuals to consistently defined local places across census years. In this paper, we describe and release public crosswalks linking the vast majority of 1790–1940 census respondents to longitudes and latitudes. These crosswalks allow analysts to explore sub-county research questions and trends using public census data.

We present two applications of our crosswalks to demonstrate their value. First, we iteratively cluster geocoded places with close neighbors, producing a consistent definition of place. This application addresses the common concern that regularly-updated county borders and shifting municipal boundaries make it difficult to match and compare places over time. Second, we test the predictions of Zipf's Law and Gibrat's Law in a historical context, finding clear deviations from theoretical predictions about the city size and city growth distribution. While these findings have been observed in more aggregated data, we are the first to illustrate these patterns over long time periods with national data on "places."

The most important contribution of this paper is the method itself. Researchers spend significant and often-duplicated time standardizing common spatial datasets, particularly when working with historical sources. We hope that the process we present in this paper will help researchers link other datasets that include unstructured place names—for example, patent data or birth and death records—to sub-county locations.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.eeh.2022.101477.

[18] The results are unchanged if we include the largest cities, but including the largest cities in the panels makes it more difficult to see the non-monotonic population growth rate dynamics for cities with lower initial population levels in 1870.

References

Aaronson, Daniel, Hartley, Daniel, Bhashkar, Mazumder, 2021. The effects of the 1930s HOLC“ redlining” maps. Am. Econ. J. Econ. Policy 13 (4), 355–392.

Aaronson, D., Hartley, D., Mazumder, B., 2021. The effects of the 1930s HOLC redlining maps. Am. Econ. J. 13 (4), 355–392.

Aaronson, D., Mazumder, B., 2011. The impact of Rosenwald schools on black achievement. J. Polit. Economy 119 (5), 821–888.

Abramitzky, R., Boustan, L., Rashid, M., 2020. Census linking project: version 1.0 [dataset]. Data retrieved from, https://censuslinkingproject.org.

Berkes, E., Karger, E., Nencka, P., 2022a. The Census Place Project: A Method for Geolocating Unstructured Place Names; Crosswalks. Ann Arbor, MI. Inter-University Consortium for Political and Social Research [distributor]. 10.3886/E179401V1

Berkes, E., Karger, E., Nencka, P., 2022b. The census place project: a method for geolocating unstructured place names; replication code. https://github.com/ezrakarger/census_place_project.

Berkes, E., Nencka, P., 2021. Knowledge access: the effects of carnegie libraries on innovation.

Brooks, L., Lutz, B., 2019. Vestiges of transit: urban persistence at a microscale. Rev. Econ. Stat. 101 (3), 385–399.

Connolly, K., 2021. How does access to college affect long-term life outcomes? Evidence from U.S. openings of two-year public colleges.

Desmet, K., Rappaport, J., 2017. The settlement of the United States, 1800–2000: the long transition towards Gibrat’s law. J. Urban Econ. 98, 50–68.

Donaldson, D., Hornbeck, R., 2016. Railroads and american economic growth: a “market access” approach. Q. J. Econ. 131 (2), 799–858.

Feigenbaum, J., Gross, D.P., 2021. Organizational Frictions and Increasing Returns to Automation: Lessons from AT&T in the Twentieth Century. Available at SSRN 3912116.

Ferrara, A., Testa, P., Zhou, L., 2022. New Area-and Population-based Geographic Crosswalks for US Counties and Congressional Districts, 1790–2020. Available at SSRN 4019521.

Fishback, P.V., LaVoice, J., Shertzer, A., Walsh, R., 2020. Race, Risk, and the Emergence of Federal Redlining. No. w28146. National Bureau of Economic Research.

Gabaix, X., 1999. Zipf’s law for cities: an explanation. Q. J. Econ. 114 (3), 739–767.

Giesen, K., Südekum, J., 2011. Zipf’s law for cities in the regions and the country. J. Econ. Geogr. 11 (4), 667–686.

Hanlon, W.W., Heblich, S., 2021. History and Urban Economics. no. 27850. National Bureau of Economic Research. accepted at Regional Science and Urban Economics

Hodgson, C., 2018. The effect of transport infrastructure on the location of economic activity: railroads and post offices in the American West. J. Urban Econ. 104, 59–76.

Hornbeck, R., 2010. Barbed wire: property rights and agricultural development. Q. J. Econ. 125 (2), 767–810.

Ioannides, Y.M., Overman, H.G., 2003. Zipf’s law for cities: an empirical examination. Reg. Sci. Urban Econ. 33.2, 127–137.

Karger, E., 2021. The long-run effect of public libraries on children: evidence from the early 1900s.

Logan, J.R., Jindrich, J., Shin, H., Zhang, W., 2011. Mapping America in 1880: the urban transition historical GIS project. Hist. Methods 44 (1), 49–60.

Lonich, D.W., 1993. Metropolitanism and the Genesis of Municipal Anxiety in Allegheny County. Western Pennsylvania History: 1918–2018, pp. 79–88.

Manson, S., Schroeder, J., Riper, D.V., Kugler, T., Ruggles, S., 2021. IPUMS National Historical Geographic Information System: Version 16.0 [dataset]. IPUMS, Minneapolis, MN doi:10.18128/D050.V16.0.

Michaels, G., Rauch, F., Redding, S.J., 2012. Urbanization and structural transformation. Q. J. Econ. 127 (2), 535–586.

Nagy, D. K., 2020. Hinterlands, city formation and growth: evidence from the US westward expansion.

Otterstrom, S.M., Price, J.P., Leeuwen, J.V., 2021. Using linked census records to study shrinking cities in the United States from 1900 to 1940. Prof. Geogr. 1–14.

Perlman, E. R., 2014. Tools for harmonizing county boundaries [computer software]. Retrieved from http://people.bu.edu/perlmane/code.html.

Ruggles, S., Flood, S., Foster, S., Goeken, R., Pacas, J., Schouweiler, M., Sobek, M., 2021. IPUMS USA: Version 11.0 [dataset]. IPUMS, Minneapolis, MN doi:10.18128/D010.V11.0.

Shertzer, A., Walsh, R.P., Logan, J.R., 2016. Segregation and neighborhood change in northern cities: new historical GIS data from 1900–1930. Hist. Methods 49 (4), 187–197.

United States Geological Survey, 2021. Geographic names information system. Updated August 27, Accessed at https://www.usgs.gov/core-science-systems/ngp/board-on-geographic-names/download-gnis-data.