Routledge
Taylor & Francis Group

Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/vhim20

New area- and population-based geographic crosswalks for U.S. counties and congressional districts, 1790–2020

Andreas Ferrara, Patrick A. Testa & Liyang Zhou

To cite this article: Andreas Ferrara, Patrick A. Testa & Liyang Zhou (16 Jul 2024): New area- and population-based geographic crosswalks for U.S. counties and congressional districts, 1790–2020 , Historical Methods: A Journal of Quantitative and Interdisciplinary History, DOI: 10.1080/01615440.2024.2369230

To link to this article: https://doi.org/10.1080/01615440.2024.2369230

View supplementary material

Published online: 16 Jul 2024.

Submit your article to this journal

View related articles

View Crossmark data

HISTORICAL METHODS: A JOURNAL OF QUANTITATIVE AND INTERDISCIPLINARY HISTORY
https://doi.org/10.1080/01615440.2024.2369230

New area- and population-based geographic crosswalks for U.S. counties and congressional districts, 1790–2020*

Andreas Ferrara[a,c], Patrick A. Testa[b] and Liyang Zhou[a]

[a]Department of Economics, University of Pittsburgh, Pittsburgh, PA, USA; [b]Department of Economics and the Murphy Institute, Tulane University, New Orleans, LA, USA; [c]National Bureau of Economic Research, Cambridge, MA, USA

ABSTRACT

In applied historical research, geographic units often differ in level of aggregation across datasets. One solution is to use crosswalks that associate factors located within one geographic unit to another, based on their relative *areas*. We develop an alternative approach based on relative *populations*, which accounts for heterogeneities in urbanization within counties. We construct population-based crosswalks for 1790 through 2020, which map county-level data across U.S. censuses, as well as from counties to congressional districts. Using official census data for congressional districts, we show that population-based weights outperform area-based ones in terms of similarity to official data.

KEYWORDS

Historical research; areal interpolation; boundary harmonization; geographic crosswalks; spatial distribution of economic activity

JEL CODES

R12; C18; C59

1. Introduction

Social scientists frequently analyze data with a geospatial component.[1] In doing so, datasets associated with different levels of spatial aggregation often need to be merged—for instance, when trying to combine county- and commuting-zone-level variables (e.g., Autor et al. 2020). The boundaries of geographic units may also change over time, as with U.S. counties across census years. If individual-level or other highly local data are not available, researchers must rely on *crosswalks* to associate aggregate data across different units. Crosswalks serve to map data associated with some *origin* spatial unit to the boundaries of the *reference* unit on which the analysis focuses.

A common approach to this boundary harmonization process involves mapping an origin unit's factors to different reference units based on the relative *areas* of overlap (Markoff and Shapiro 1973; Goodchild, Siu, and Lam 1980; Hornbeck 2010). As an illustative example, suppose a researcher wished to associate data for U.S. counties as of 1880 with county boundaries as of 1870, in the interest of having consistent spatial units over time. Importantly, some county boundaries changed between 1870 and 1880 and thus do not coincide across these years. For instance, suppose county C^{70} split after 1870 into two counties: C_1^{80}, which lies totally in C^{70}, and C_2^{80}, which lies only partly in C^{70}. To approximate the portion of C_2^{80}'s factors that exist within C^{70}'s boundaries, the researcher could intersect both sets of boundaries and compute the share of C_2^{80}'s area, $a<1$, that lies in C^{70}. Then, the researcher could re-aggregate each factor of interest within C^{70} by taking the weighted sum of the values from C_1^{80} and C_2^{80}, using the area shares computed in the previous step as weights (i.e., 1 and a, respectively).[2] A core assumption underlying this procedure is that the factors measured by the aggregate data (e.g., population stocks) are *uniformly distributed* in space within the boundaries of the origin units being disaggregated. A large set of papers has adopted this approach for the purposes of both intertemporal spatial analysis (see Hornbeck and Naidu 2014; Lee and Lin 2018; Bazzi, Fiszbein, and Gebresilasse 2020; Calderon, Fouka, and Tabellini 2023; Ferrara and Testa 2023; Han, Milner, and Mitchener 2023) and spatial harmonization across different contemporaneous units (Eckert et al. 2020; Testa 2021; Bazzi et al. 2023).

This paper makes four contributions to this body of work, with potential for broad application among economic historians, urban economists, political

CONTACT Andreas Ferrara ✉ a.ferrara@pitt.edu Department of Economics and NBER, University of Pittsburgh, Pittsburgh, PA, USA.

*Thanks to Rick Hornbeck, Allison Shertzer, and Sarah Walker for helpful comments and suggestions. The area- and population-based crosswalks produced in this paper, teaching material, as well as code and data for the replication exercise can be downloaded at https://doi.org/10.3886/E150101. All errors are our own.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/01615440.2024.2369230.

scientists, and other spatial researchers. First, we address prevailing concerns that the uniformity assumption underlying area-based weights may generate errors in harmonized data, to the extent that boundaries do not neatly coincide across origin and reference units (Gregory 2002; Logan, Stults, and Xu 2016; Hanlon and Heblich 2022). To do this, we apply a procedure for generating a set of *population-based* weights in the context of the conterminous U.S. between 1790 and 2020, based on several spatial models of historical sub-county population distribution (Fang and Jawitz 2018; Leyk and Uhl 2018). We use these data to produce crosswalks that relax the spatial uniformity assumption and identify where populations are more concentrated within counties. This is useful for cases in which boundary harmonization involves spatial disaggregation of county-level stock data. In such cases, identifying where people disproportionately live within a county lets us assign larger weights to data for some parts of counties than their areal coverage might entail under an area-based approach. This is particularly important for data that are likely to be correlated with population density, such as total income and the number of college-educated workers.

Second, we use these new weights to extend previous county-to-county crosswalks across all U.S. census years (Hornbeck 2010; Eckert et al. 2020). These build algorithmically on previous approaches in Schroeder (2016), who models historical population distributions within 2010 U.S. county boundaries, and Beddow and Pardey (2015), who use information on the spatial distribution of production in the U.S. as of 2000 to map historical county-level crop data to those boundaries. Our resource is complementary to the work of Berkes, Karger, and Nencka (2023)—whose approach granularly geocodes individuals to towns and cities for the 1790–1940 U.S. Censuses—for cases in which sub-county data are not available to the researcher.

Third, we use both area- and population-based models to generate a novel database of county-to-congressional district (CD) crosswalks for the entirety of U.S. history. An expansive set of research in political science and historical political economy entails analysis at the CD level (e.g. Lee, Moretti, and Butler 2004). Yet, relevant aggregate data are much more likely to be available at the county level, whose boundaries often do not coincide neatly with CD boundaries. Meanwhile, fully disaggregated data seldom associate individuals with their CD. CDs also offer a particularly relevant application of our population-based weights: to the extent that more densely-populated areas are associated with smaller CDs, area-based weights are likely to underestimate the populations of an urban CD and overestimate the populations of a non-urban CD located within the same county. The more concentrated urban agglomeration is relative to a county's area (e.g., as in mountainous or marshland areas), the greater this bias is likely to be. Population-based weights help us overcome such bias.

Lastly, we provide a formal test of the performance of area- and population-based crosswalks, by comparing data that were collected at the CD level with those generated from crosswalked county-level information. For this purpose, we replicate the CD-level data and key estimates in Lee, Moretti, and Butler (2004). To measure CD characteristics, the authors importantly use official CD-level data from the U.S. Census of Population and Housing for 1960 through 1990. These ground-truth data allow us to evaluate the performance of the area- versus population-based weighting approach when crosswalking county-to-CD level aggregates. Using county-level census data from Haines (2010), we show that while both area- and population-based crosswalks produce similar data to official measures, replicating key results in Lee, Moretti, and Butler (2004), data constructed using population-based weights consistently outperform area-based ones in terms of similarity to official measures. In particular, the average accuracy of the data constructed with the population-based crosswalks is almost 20% higher than those using the area-based data. The best-performing crosswalk in this application uses built-up property data to construct population-based weights. We conclude by discussing some limitations of population- and area-based crosswalks. All crosswalks, teaching material, and replication files can be downloaded from https://doi.org/10.3886/E150101.

2. Constructing and implementing geographic crosswalks

In this section, we describe the methods for constructing and then using our area- and population-based crosswalks, with the intention of providing applied researchers with prerequisite background knowledge and intuition for using the crosswalks. We will focus primarily on our county-to-congressional-district (CD) crosswalks, which span the 1st through 116th U.S. Congresses from 1790–2020, as harmonization across geospatial units defined at different levels of aggregation is particularly prone to the problems being addressed in this paper. These methods generalize to the harmonization of county boundaries across U.S.

censuses.[3] For the construction of these crosswalks, we use data for county boundaries provided by Manson et al. (2020) and CD boundaries from Lewis et al. (2021).

We generate full three sets of county-to-CD crosswalks, based on: (i) the nearest census year, relative to the starting year of a given Congress; (ii) the census decade shared with the starting year of a given Congress; and (iii) the census of apportionment associated with a given Congress.[4] This is to provide researchers with sufficient flexibility to choose the time dimension that best suits their application. Each of these sets includes six kinds of weights:

1. Area-based (model 1, or M1).
2. Population-based (M2), with county area divided into urban and rural areas.
3. Population-based (M3), with county area divided into urban and rural areas after excluding non-inhabitable areas.
4. Population-based (M4), with county area divided into urban and rural areas after excluding non-inhabitable areas, with additional weighting for topographic suitability (i.e., elevation).
5. Population-based (M5), with built-up settlement areas indicated in space (1810–2020 only).
6. Population-based (M6), with built-up property counts indicated in space (1810–2020 only).

M1 is equivalent in construction to existing area-based crosswalks. M2–M4 use maps based on historical population estimates for 1 × 1 kilometer grid cells from Fang and Jawitz (2018), whereas M5–M6 use maps based on historical property records for 250 × 250 meter grid cells from Leyk and Uhl (2018).

In addition to these county-to-CD crosswalks, we also construct *county-to-county* crosswalks for all pairs of censuses from 1790 to 2020, using both area-based weights and our population-based weights. Between these county-to-CD and county-to-county crosswalks, our crosswalks can be used to harmonize boundaries of any county to any CD in U.S. history for all incorporated conterminous U.S. states.

2.1. Area-based harmonization

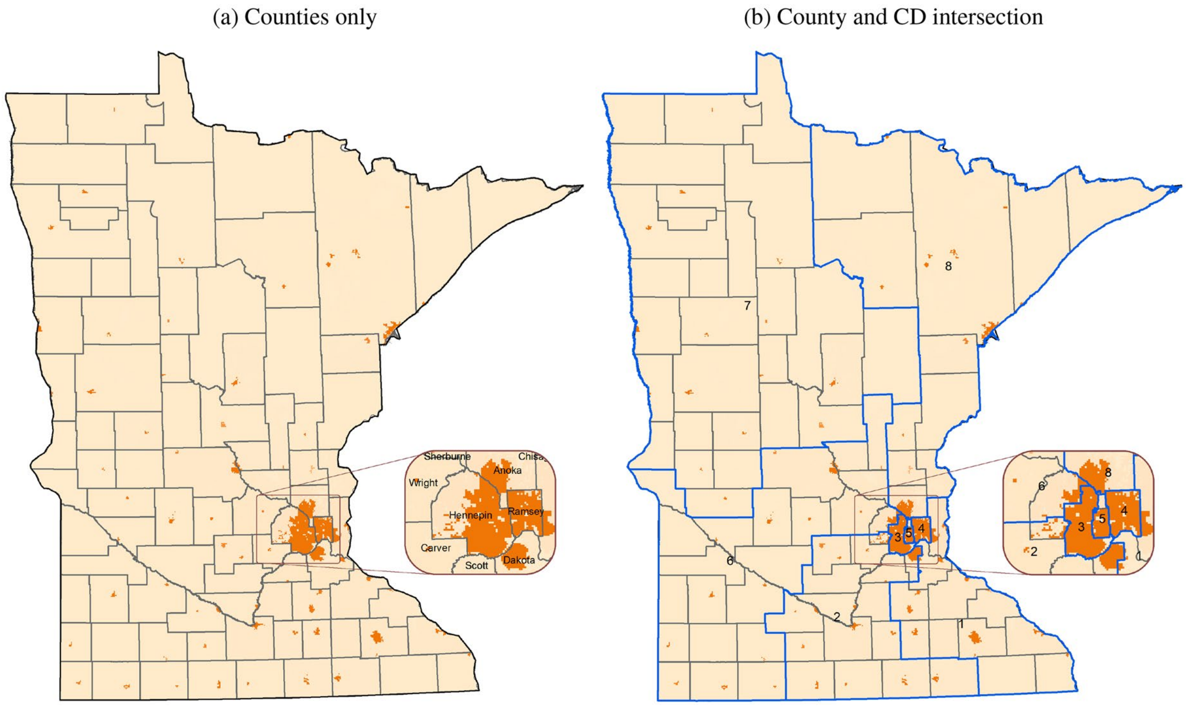
Area-based approaches to boundary harmonization procedures generally entail a process of spatial disaggregation and re-aggregation. For the county-to-CD case, this process involves intersecting a map of counties from a particular census year with a map of CDs from a given Congress year. Counties are then disaggregated into a set of sub-county units (henceforth "county-parts"), based on the CD in which they are located. For example, counties that are intersected by a single CD boundary are located partly in two CDs and thus have two county-parts. Meanwhile, counties that lie wholly within a CD without intersecting its boundaries are their own and only county-part. We then calculate the areas (in square meters) of all counties, all CDs, and all county-parts.[5] Once counties are disaggregated given their intersections, county-parts can then be re-aggregated based on their associated CD, with the sum of the areas of the county-parts matching the area of the whole CD.

In the process, the data values associated with the initial counties (e.g., total population, total number of Blacks) may be re-associated with CDs. Under an area-based procedure, each county-part is assigned each of its county's data values, which are then *weighted* by the share of the county's total area that lies in that county-part. These county-parts and their weights, which add up to 1 for each county undergoing harmonization, comprise the crosswalk under a strictly area-based approach (i.e., M1 above). Using these, a given CD's data values are in turn approximated using the aggregates of weighted values, summed across all counties that have a county-part located within that CD. Values associated with a county whose area is shared equally by two CDs are each weighted by 0.5, while values associated with a county that lies wholly within a CD are weighted by 1. In the Online Appendix, we describe the process and data used to generate these weights in ArcMap for a given census and Congress year pair.

2.1.1. Example: Minnesota

Minnesota offers a useful case study of this method. Figure 1 shows Minnesota's county boundaries in 1970 and its congressional district boundaries as of 1973. Note that CD 7, in the state's northwest corner, consists only of whole counties. We can add up the values of each stock variable across these 27 counties within CD 7, and it will give us CD 7's values for those same variables. The same goes for CD 4, which consists only of Ramsey County. If every county in Minnesota had a population of 1,000 in 1973, CD 7 would have 27,000 residents, while CD 4 would have 1,000.

Other counties, like CD 8 in the state's northeast corner, may consist of whole counties and/or portions of other counties. For CD 8, an example of the latter case is Anoka County, of which a small



(a) Counties only

(b) County and CD intersection

Figure 1. Minnesota counties, CDs, and population distribution based on 1970 U.S. Census.
Note: This figure shows the land area of the state of Minnesota with population distribution information for 1970, where darker orange implies a greater number of residents per square kilometer. The gray boundaries show the state's county boundaries as of the 1970 U.S. Census. The thicker, blue lines in panel (b) show the state's congressional district (CD) boundaries as of the 93rd Congress (1973-4). County shapefiles are from Manson et al. (2020). CD shapefiles are from Lewis et al. (2021). Population distribution information for 1970 comes from M3 in Fang and Jawitz (2018).

portion—about 1/20th of the area of the whole county—is instead part of CD 5 together with part of Hennepin County. Hence, under an area-based crosswalk, 19/20th of the population and of other stock variables associated with Anoka County are associated with CD 8. If every county in Minnesota had a population of 1,000, CD 8, which also consists of 10 whole counties, would be estimated as having 10,950 residents.

There are potential drawbacks to using this area-based method when origin and reference unit boundaries do not neatly coincide, as in the latter case. Chiefly, the output is accurate only under certain conditions on the distribution of population. To see this, note the background coloration of Figure 1, which plots alongside county and CD boundaries a map of population distribution from Fang and Jawitz (2018). This shows that, although only about a tenth of the area of Hennepin County is located within CD 5, the part that *is* includes some of the most populated areas of the county (as shown in dark orange). Yet despite the fact that this part of Hennepin County is among the most densely populated areas in the county, an area-based approach would assign only 10% of the county's population to CD 5—significantly underweighting this county-part, while overweighting all of the others.

2.1.2. When is an area-based crosswalk appropriate?

Suppose a researcher is attempting to associate several county-level stock variables with congressional districts. For the area-based weights to be appropriate in settings where county and CD boundaries overlap, the following *uniformity* condition is key:

Assumption (Uniformity). *Let C be any continuous, two-dimensional county with area $c>0$ and a vector of positive and finite values $P=(p_1,p_2,\ldots,p_n)$. Let A be any continuous, two-dimensional subset of C with area $ac\in(0,c)$ and a vector of positive and finite values $R=(r_1,r_2,\ldots,r_n)$. C satisfies uniformity in population distribution if $R=aP$ for all $A\subset C$.*

In this definition, P and R represent the set of stock variables at the county and sub-county (e.g., neighborhood) level, respectively, such as total population, total income, the total number of Spanish-speakers, etc. in their respective areas. Hence, when uniformity holds, a neighborhood's share of a given sub-population in a county is always equal to its share of the county's total area.[6] Under this condition, an area-based crosswalk would accurately map data associated with one set of spatial units (e.g., counties) to the boundaries of another (e.g., congressional districts). This might be plausible in relatively low-density settings, such as farmland, with spatially homogeneous populations; when harmonization involves highly disaggregated data; or when the "origin" units being harmonized lie neatly within the "reference" units on which the analysis focuses, with little overlap in boundaries. For example, a researcher studying a sample of U.S. counties across several decades may be able to re-aggregate counties backward in time, as in Hornbeck (2010).

In many settings, however, uniformity will not hold, e.g., due to the presence of agglomeration forces making the distribution of population uneven across space. In such cases, area-based crosswalks will generate errors relative to ground-truth data whenever origin units must be disaggregated, such as when a county lies in two or more CDs. The more often the boundaries of origin and reference units do not coincide, the more such error will occur and accrue. To address this concern, we construct a set of population-based crosswalks in addition to the area-based crosswalk, which allow for heterogeneous population distributions within counties. We then compare the relative performance of these crosswalks.

2.2. Constructing population-based crosswalks

We now seek to relax the uniformity assumption, through the use of information on historical sub-county population distribution from Fang and Jawitz (2018) (for short, FJ) and Leyk and Uhl (2018) (for short, LU). FJ estimate historical population counts for 1 × 1 kilometer grid cells, which we use to construct a set of population-based weights. These include: (i) model 2 (M2), which is based on a division of counties into urban and rural areas, with urban population counts being distributed around city centers according to the power law scaling relationship detailed below;[7] (ii) model 3 (M3), which is is based on a version of M2 that first excludes non-inhabitable areas, such as bodies of water or areas where settlement is legally restricted, such as national or state park; and (iii) model 4 (M4), which is based on a version of M3 that also weights population counts based on topographic suitability as measured by county mean elevation. LU, in contrast, derive proxies for historical population size for more granular 250 × 250 meter grid cells based on historical property records data, which they show to be highly correlated with local population size. We use these to construct two further weights: (i) model 5 (M5), which is based on their binary measure of "built-up area," which assigns a value of 1 to a grid cell if it contains at least one built-up property record in a given year, and (ii) model 6 (M6), which is based on the "built-up property" counts themselves, summing the number of records (e.g., building units) within the grid cell in a given year.[8] We will now describe the underlying spatial models from FJ and LU in greater depth, after which we will discuss how we use these to construct the crosswalk weights themselves.

2.2.1. Describing the spatial models in Fang and Jawitz (2018)

Given that historical sub-county spatial data for population hardly exist, FJ's models first estimate the spatial extent of urban areas for the conterminous US over time using population distribution information for urban areas from the 2000 U.S. Census. Concretely, FJ extrapolate the size of the urban area to previous census years, using the following power law scaling relationship,

$$A_{U,\varphi} = \alpha_\delta P_{U,\varphi}^{\beta_\delta} \tag{1}$$

where $P_{U,\varphi}$ is the population size of urban area φ in U.S. Census division δ in a given year, and where α_δ and β_δ are the coefficients of the power function, which are fixed scaling factors based on the areas and populations of U.S. cities in 2000. Using historical population data from the census, FJ then estimate the historical areal extents of urban areas back to 1790, within which population counts are distributed according to the models described above.

The motivation for the use of such a power law distribution comes from Chen (2015) and has famously found applications in describing other urban regularities, such as Zipf's law. Generally, the growth and size of urban areas has been shown to follow remarkably robust statistical distributions (see Eeckhout 2004), and even large scale shocks tend to not alter cities' population growth trajectories over the long-run (Davis and Weinstein 2002; Miguel and Roland 2011). All of FJ's models of sub-county population rely on this assumption, while a subset make further adjustments for the presence of non-inhabitable areas and topographic suitabilities—the basis for our weights M3 and M4, respectively. For a more in-depth description of these models and the data used to construct them, see the Online Appendix. For additional discussion of FJ's assumptions and potential drawbacks, see Section 4.

2.2.2. Describing the spatial models in Leyk and Uhl (2018)

In contrast to FJ, LU derive maps of historical urban settlements from property records data in the Zillow

Table 1. Correlation, root MSE, and MAE between CD-level data and county-level data crosswalked to the CD-level using area- and population-weighting.

	Population			Income			Urban population		
	(σ = 379,399.9)			(σ = 10,642.2)			(σ = 311,054.5)		
Model	Corr.	RMSE	MAE	Corr.	RMSE	MAE	Corr.	RMSE	MAE
M1	0.827	259,040.1	120,895.3	0.724	10,509.2	4,956.2	0.732	266,936.9	133,663.4
M2	0.961	110,015.7	55,366.1	0.910	5,019.8	2766.8	0.926	122,646.4	71,128.9
M3	0.957	114,766.7	57,696.0	0.908	5,112.1	2,830.2	0.921	127,017.5	73,332.6
M4	0.961	109,053.6	53,777.6	0.913	4,937.3	2,724.8	0.928	121,398.3	69,432.3
M5	0.946	133,773.2	75,240.5	0.894	5,637.2	3,192.3	0.903	145,714.3	88,863.6
M6	0.978	83,491.9	34,670.4	0.931	4,399.6	2,256.3	0.959	93,615.3	49,773.6
	Black population			Manufacturing employment			Voting population		
	(σ = 83,206.6)			(σ = 46,539.5)			(σ = 229,314.4)		
Model	Corr.	RMSE	MAE	Corr.	RMSE	MAE	Corr.	RMSE	MAE
M1	0.624	68,357.9	30,660.0	0.856	28,001.3	12,276.6	0.785	180,758.4	86,966.0
M2	0.715	58,798.8	26,211.6	0.972	11,410.5	6,084.0	0.946	79,386.8	42,107.0
M3	0.704	59,888.9	26,634.2	0.969	11,992.5	6,304.4	0.942	82,445.3	43,722.8
M4	0.711	59,300.7	26,114.6	0.972	11,340.0	5,929.0	0.947	78,449.2	41,046.8
M5	0.696	62,006.3	29,549.6	0.957	14,557.7	7,938.8	0.929	94,659.7	55,178.0
M6	0.752	56,392.5	26,050.0	0.985	8,585.6	4,678.6	0.971	58,353.2	27,093.2

Note: This table compares harmonized CD-level data generated by our six crosswalk weights to official CD-level census data, as featured in Lee, Moretti, and Butler (2004), from the U.S. Census of Population and Housing of 1960, 1970, 1980, and 1990. We report correlations (as shown visually in Figure 3) together with the root mean squared error (RMSE) and the mean absolute error (MAE). We also report the standard deviation (σ) of the underlying CD-level variable. See the notes to Figure 3 for other details.

Transaction and Assessment Database (ZTRAX) beginning in 1810. Records of building and building units are mapped to 250 × 250 meter grid cells, which can then be aggregated within county or other polygons. Comparisons with county-level population data for 1860–2010 in their Table 1 show that a one unit increase in built-up property records within a county is associated on average with 2.68 (0.01) additional residents, with these records accounting for nearly 93% of the variation in total population size over time across sample counties. Property records thus potentially provide an accurate and granular proxy for historical population counts. Based on these property records, LU construct several maps, including ones based on a binary measure of "built-up areas" and another based on "built-up property" counts themselves—the basis for our weights M5 and M6, respectively. For more descriptions of the LU models and their underlying data, see the Online Appendix. Sections 3 and 4 further discuss the different models and their performance.

2.2.3. Constructing the crosswalks

In order to relax the uniformity assumption, our population-based crosswalks base the disaggregation of county-level data no longer on relative area but rather on relative *population*, using the models of historical sub-county population distribution from FJ and LU. The resultant maps allow us to calculate for each census year a total population count (or property-based proxy) for each county, P_C, as well as for each county-part within that county that lies in a different CD, P_A, with $\sum_{A\in C} P_A = P_C$. These values are calculated by summing the grid cell values within those respective polygons, using GIS software. Then, similar to the area-based crosswalk, we use the ratio of P_A to P_C as a weight for each county-part, with which to ultimately multiply a county's relevant stock data prior to its aggregation to the CD level.[9]

In contrast to the area-based crosswalk, relatively small county-parts in terms of area might in some cases receive a relatively *large* crosswalk weight—for instance if they are associated with an urban area. Such discrepancies between area- and population-based weights are shown in Figure 2, which relates weights from each of the five population-based models to those from the area-based one for the 2010 U.S. Census and the 112th Congress. Although weights are highly correlated across models overall, many individual weights differ significantly.

For example, take Dorchester County, SC, a suburban county that partly overlaps with the Charleston metropolitan area. As of 2011, nearly 80% of its area lied within CD 6. At the same time, around 80% of its population instead lived in the much smaller and more urban CD 1. M1 would have associated around 80,000 Dorchester residents with the wrong congressional district during the harmonization process, something remedied by the population-based models.

Even more extreme is Monroe County, FL. Over 99% of its residents live in the very tiny Florida Keys, represented in 2011 by CD 18, whereas around 85%

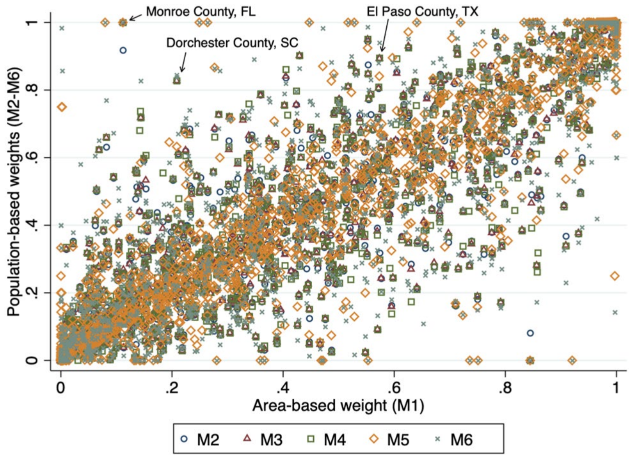


Figure 2. Comparison of area- and population-based weights. *Note:* Figure shows the relationship between our area-based weights and each of our population-based weights for 7,493 county-parts, based on 3,109 counties from the 2010 U.S. Census and 432 congressional districts (CDs) from the 112th Congress (2011-12). These exclude Alaska and Hawaii, for which Fang and Jawitz (2018) and Leyk and Uhl (2018) lack historical population distribution information.

of its area, mostly wetlands, were in CD 25. The more concentrated the urban area relative to the size of the county, the more likely these discrepancies are to exist, as they do in desert areas like Phoenix, AZ, and Las Vegas, NV, as well as in swamp and wetland areas like Southern Louisiana and the Florida Peninsula.

2.3. Implementing our crosswalks

Our crosswalks can be used to harmonize historical county boundaries to those from any other census period between 1790 and 2020. Whether using area- or population-based crosswalks, a possibly crucial choice is which base year to pick. A commonly-used option is to crosswalk to the year in which units have the largest spatial extent—with the caveat that this may generate some loss of spatial precision due to sample aggregation.[10] Our crosswalks can also be used to harmonize county boundaries to contemporaneous CD boundaries. These include three options, based on counties associated with: (i) the nearest census year, relative to the starting year of a given Congress; (ii) the census decade shared with the starting year of a given Congress; and (iii) the census of apportionment associated with a given Congress. Each crosswalk file includes weights from M1–M6, except for 1790 and 1800, which include only M1–M4, and except for Alaska and Hawaii, which have weights based on M1 only. Between the county-to-CD and county-to-county crosswalks, our crosswalks can be used to harmonize the boundaries of any county to any CD in U.S. history for all incorporated conterminous U.S. states.

The process of implementing these crosswalks is straightforward. We will illustrate this process using an example. Suppose one were interested in harmonizing data defined for 1960 U.S. county boundaries to CD boundaries for the 88th Congress. Suppose the variable of interest is the percent of the population that was born in Mexico. One would do the following:

1. Obtain the county-level data for 1960 for two variables: (i) total population and (ii) total number of persons born in Mexico. It is critical to harmonize only county-level stock variables for weights to be appropriate. If source data are shares or average outcomes, one should transform the variable first, e.g., by multiplying by total population.
2. Given some set of county identifiers (e.g., FIPS or NHGIS codes), merge the 1960 county file with the 1960 to 88th Congress crosswalk file. This expands the set of counties into the full set of county-parts, based on the CDs they are associated with.
3. Take note of which counties are not merged successfully or contain missing data. In the latter case, data for the CDs in which they lie should likely be considered missing as well. Then multiply the stock variables by the weights associated with the county-parts. Weights may differ across the six models in our crosswalk.
4. Finally, collapse (i.e., sum) the weighted counts for each variable by CD identifiers. Round or mark as missing any cell as needed. The unit of observation is now the CD.

See the Online Appendix for sample Stata and R code demonstrating this process.

3. Application

In this section, we showcase the usefulness and accuracy of our county-to-CD crosswalks, by replicating the CD-level data and the balance tests that underscore the regression discontinuity (RD) empirical strategy used in Lee, Moretti, and Butler (2004). RD designs exploit plausibly-random variation in exposure to some treatment, by comparing groups just below and above the treatment's intervention threshold. In empirical political economy, one application of RD considers the effects of elections based on partisan

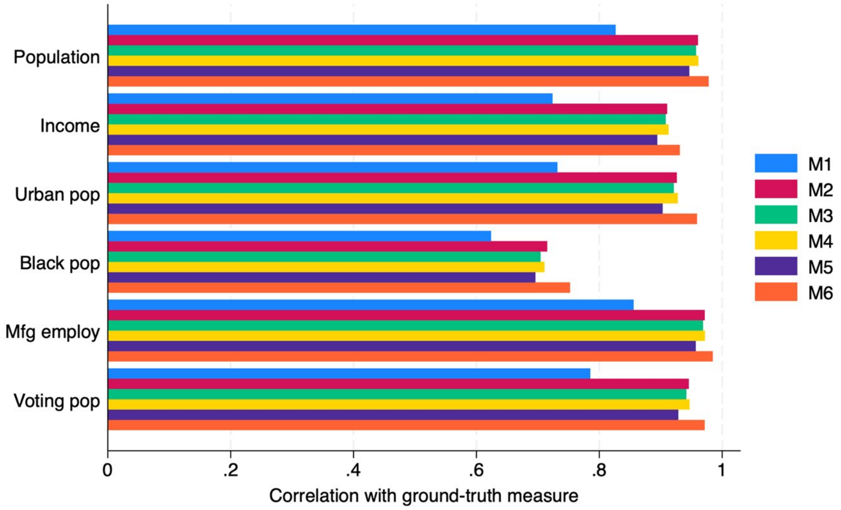


Figure 3. Comparing harmonized data with official CD-level census data.
Note: Figure compares harmonized CD-level data generated by our six crosswalk weights to official CD-level census data, as featured in Lee, Moretti, and Butler (2004), from the U.S. Census of Population and Housing of 1960, 1970, 1980, and 1990. These are defined for CD boundaries for the U.S. Congresses at the top of the corresponding apportionment periods—the 88th, 93rd, 98th, and 103rd U.S. Congresses, respectively. These boundaries are assumed fixed for each decade in Lee, Moretti, and Butler (2004). We therefore limit our comparisons here to those four Congresses, for which the official data correspond to the true measures for each district. One advantage to our crosswalks is that they can harmonize county-level data to CD boundaries for *any* Congress, allowing researchers to account for changes in CD boundaries between congressional apportionments.

representation, comparing places (e.g., CDs) where a party's candidates narrowly won against those where they narrowly lost. In principle, such places around this "tied-election" threshold are likely to be highly similar. In practice, balance tests, such as those used in Lee, Moretti, and Butler (2004), serve to test for the exogeneity of the CD characteristics around the tied-election threshold in support of this identification strategy. To measure CD characteristics, the authors importantly use the official CD-level data from the U.S. Census of Population and Housing for 1960 through 1990. We use county-level census data from Haines (2010) to test whether these data, and in turn the results of the balance tests in Lee, Moretti, and Butler (2004), are replicated when CD characteristic data are harmonized from county-level data, as well as whether this differs across our six crosswalk weighting models.

We begin by using our crosswalks to construct CD-level stock data from the county-level census data, with which we compare to the official CD-level census data used in Lee, Moretti, and Butler (2004). We focus on six variables for which we can confidently reconstruct the data: (i) total population, (ii) total real income, (iii) urban populations (of 2,500+ inhabitants), (iv) Black population, (v) number of manufacturing workers, and (vi) number of eligible voters.[11] These reconstructions compare favorably to the official CD-level census data across all five of our population-based models, as gauged by their correlations with the former as shown in Figure 3. On average, the correlations between the data values generated using M2–M6 and the official data are around 0.91, whereas the correlation is about 0.76 for M1. This means that our population-based approach improves the correlation with the official data by almost 20% relative to an area-based one. This result mirrors the evaluation of area-based interpolation for 2000-10 census tract data in Logan, Stults, and Xu (2016), who show that such approaches can lead to large errors. Meanwhile, among the five population-based models, none clearly or consistently outperform the others, with the exception of M6—though we will discuss the limitations and inherent tradeoffs of using the various models more in Section 4.

Table 1 reports further summary statistics alongside these correlations, which yield similar implications. Note that the standard deviation in population in the CD-level census data featured in Figure 3 is about 379,400 individuals. Relative to this value, the root MSE corresponding to M1 is about 259,040 individuals in terms of the deviation from the census-based measure, which is 68% of a standard deviation. For M6, the root MSE is 83,492 individuals, which is only 22% of a standard deviation.

At the same time, harmonization is more successful for some variables than others, regardless of the model used. Of the six variables we reconstruct, the total population and manufacturing population data are closest to the official CD-level census data, while the

Table 2. LMB's Balance tests using congressional district-level data versus harmonized CD data constructed from county-level information.

	Difference in district population between democrat and republican districts					
	(1)	(2)	(3)	(4)	(5)	(6)
Total pop (M1)	−92,262.9***	−72,968.3***	−23,473.9**	−24,417.8*	−34,212.2	−12,495.7
	(12,117.1)	(12,874.6)	(11,741.6)	(12,977.4)	(22,510.6)	(21,968.6)
Total pop (M2)	−37,081.3***	−18,546.0***	−3,286.6	−3,727.6	−1,255.2	−336.5
	(5,975.7)	(5,935.8)	(6,254.6)	(7,972.6)	(12,973.8)	(13,872.5)
Total pop (M3)	−38,286.8***	−19,051.5***	−3,706.1	−4,059.8	−1,240.9	13.9
	(6,224.5)	(6,156.6)	(6,348.2)	(8,065.7)	(13,139.5)	(14,200.1)
Total pop (M4)	−32,030.1***	−14,839.0**	−2,192.7	−3,198.2	1,262.0	3,320.0
	(5,950.8)	(5,905.5)	(5,958.1)	(7,710.2)	(12,850.0)	(13,732.5)
Total pop (M5)	−64,413.7***	−47,177.6***	−17,042.3**	−13,148.8	−17,519.3	−4,635.7
	(7,113.2)	(7,351.6)	(7,845.1)	(9,414.8)	(15,465.8)	(15,524.0)
Total pop (M6)	−20,360.6***	−7,620.1*	−2,138.8	−2,258.4	2,145.1	7,709.1
	(4,137.0)	(4,215.2)	(5,507.0)	(7,500.8)	(13,164.5)	(11,929.9)
Total pop (LMB)	−1,817.6	3,019.9	4,961.5	3,211.1	8,640.6	2,008.0
	(3,517.3)	(3,723.4)	(4,562.7)	(5,524.2)	(8,427.0)	(9,258.1)
Bandwidth	All	+/− 25	+/− 10	+/− 5	+/− 2	Polynomial
Observations	13,231	10,065	4,086	2,030	794	13,211

Note: Each row features estimates from a different harmonization model, except for row (7), which uses data and code from Lee, Moretti, and Butler (2004). Observation counts reflect those in row (7). Column (1) features the entire sample. Columns (2) through (5) limit the sample by varying bandwidths around the 50% mark. Column (6) includes a fourth order polynomial in Democratic vote share, which is interacted with the above-below 50% dummy. The unit of observation is the district-congress. Standard errors are clustered by district-decade. Significance levels are denoted by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

number of Blacks is the most different. Indeed, some subpopulations, such as racial minorities, may have spatial distributions that differ greatly from the overall population's (e.g., due to historical segregation), contributing to poorer performance in the harmonization process. In general, it is important to keep in mind when harmonizing data whether a particular variable is appropriate, given its spatial distribution relative to a county's area or overall population, as further discussed in Section 4.

On the other hand, this illustrates clear upsides to using our approach to harmonize county-level data to the CD level. Official CD-level data as used in Lee, Moretti, and Butler (2004) are only available for some decades and, even then, only for one Congress per decade (at the beginning of a new census apportionment period), despite CD boundaries often changing within states between censuses. They are also limited to a relatively small set of variables, whereas spatial researchers often deal with novel county-level data constructed from historical data not found in the census. In contrast, our approach is available for every Congress year and its associated boundaries, and it works with any data that can be associated with a U.S. county, at any point in time.

Lastly, we re-estimate the balance tests from Table 2 in Lee, Moretti, and Butler (2004). As a baseline, we first successfully replicate the balance tests using their official data and code. Estimates from their balance tests for population size are shown in row 7 of our Table 2, together with those based on our six weighting models. Among our models, estimates based on M4 and M6 are closest to the ground-truth ones, while M1 and M5 are the furthest, mirroring their respective performances in Figure 3. We further re-estimate the balance tests for other variables considered by Lee, Moretti, and Butler (2004); as in their paper, % urban and % Black show slight but statistically significant discontinuities across multiple specifications. We report these estimates in Tables A1–A6 in the Online Appendix. Overall, most observable characteristics show few differences between Democratic and Republican CDs around the tied-election threshold, in line with the original balance tests in Lee, Moretti, and Butler (2004).[12]

4. Discussion and conclusion

We now turn to some discussion, beginning with a few notes on interpretation. First, we want to emphasize that data generated from crosswalks, ours or otherwise, are necessarily imperfect, relative to ground-truth data. However, in the absence of ground-truth data for the unit of interest, researchers must often rely on crosswalks from some other "origin" unit in order to approximate them. Currently, researchers commonly use crosswalks based on area-based weighting. To the extent that the spatial distribution of origin data often varies with urban density, we argue that our population-based crosswalks constitute an important improvement over these existing practices. Second, we are by no means claiming that our population-based crosswalks are *always* preferred over other approaches. We now address some limitations of our population-based crosswalks.

4.1. Limitations of crosswalks

Error in harmonized data stems largely from the act of disaggregating the already-aggregated "origin" data. Hence, if the researcher has access to ground-truth data or can sum aggregated data within larger spatial units (e.g., backward in time for many U.S. counties) without the need for disaggregation, then *neither* area- nor population-based crosswalks should generally be used. Indeed, while population-based crosswalks generate less error than area-based crosswalks in many cases, they nonetheless entail nonzero error to the extent that they imperfectly approximate the spatial distribution of the origin data.

4.1.1. When should you use an area- versus population-based crosswalk?

If data *must* be disaggregated in the process of boundary harmonization, then our population-based crosswalks will be preferred to widely-used area-based approaches whenever origin data are spatially correlated with urban density. Conditional upon this, population-based crosswalks can nonetheless be expected to introduce error relative to ground-truth data as the absolute value of the spatial correlation between the stock data of interest and the total population decreases. If stock data are instead distributed more uniformly, then an area-based approach might in fact be preferred on those grounds.

If stock data are *more* unevenly distributed than the overall population, then a population-based approach will be preferred over an area-based approach, but its output will nonetheless be inaccurate relative to ground-truth data, as with the Black population in the exercise above. Note that if a variable is negatively correlated with population (e.g., air quality), such variables can be transformed prior to harmonization (e.g., into a measure of air pollution).

4.1.2. Which population-based weights? FJ versus LU

Suppose your data are indeed spatially correlated with urban density. Absent ground-truth data, when is a population-based crosswalk based on the FJ-based models (i.e., M2–M4) more appropriate, versus a crosswalk based on the LU-based approaches (i.e., M5–M6)? As it turns out, both sets of weights offer distinct advantages and limitations, which render each of them preferable under different circumstances.

When are M2–M4 more appropriate? It is important to keep in mind that, in constructing the population maps upon which M2–M4 are based, FJ rely on modeling assumptions which—despite being based on empirical regularities in available data—may entail some error in harmonized data. Recall that the areal extents of historical urban areas are estimated using the area-population power law scaling relationship in Eq. (1), projected backward from estimates derived using data from 2000. Further topographic suitability adjustments are made in the construction of M4, based on region-varying effects of elevation on log population density in the available data. These assumptions are likely to produce some error in the final maps and in turn our crosswalks. For a visual example of how these assumptions manifest spatially, see Appendix Figure A2. At the same time, the need for such assumptions, like the need for crosswalks themselves, stems from the nonexistence of these ground-truth data. If these ground-truth data existed, one would not need crosswalks to begin with. Alternative methods for estimating sub-county population distributions would entail similar limitations. For instance, Berkes, Karger, and Nencka (2023) place individuals within location centroids, but approximating areal extents beyond these centroids would require similar assumptions. Moreover, alternative sub-county population distribution data are available only for a subset of regions or census years. Spatial disaggregation using census tracts would cover only the twentieth century and would exclude many urban areas for most years, while the Census Place Project ends in 1940. In contrast, FJ estimate population distributions for the conterminous U.S. since 1790, offering time coverage that exceeds all alternatives.

In contrast, the LU approach used to construct M5–M6 forgoes modeling actual population distributions, instead relying on historical property records data to proxy granularly for population size and in turn construct maps of historical urban settlements. For this, no strong assumptions need be made. For analyses involving more modern spatial data, this offers a highly accurate alternative to ground-truth data (see Figure 3). Yet this approach has its own drawbacks, too. The ZTRAX property database from which the LU maps are constructed have gaps in the data and are increasingly unlikely to have property records for a given county moving further back in time. This is a result of both (i) imperfect record-keeping, especially in less developed regions, and (ii) increasingly sparsely-populated land shares, especially in the Western U.S. The first factor is likely to generate significant measurement error in early decades, especially for the count-based M6.[13] For a visual example of how property records, including missing data, may manifest cartographically, see Appendix Figure A3. Moreover, both of these factors

Table 3. Comparison of different crosswalk weighting models.

	Weighting scheme		Factors accounted for in sub-county population distributions				Data coverage	
Model	Area	Sub-county population	Non-inhabitable areas	Elevation	Historical property records (binary)	Historical property records (counts)	Grid cell size	Decades
M1	✓							1790-2020
M2		✓					1 × 1 km	1790-2020
M3		✓	✓				1 × 1 km	1790-2020
M4		✓	✓	✓			1 × 1 km	1790-2020
M5		✓	✓		✓		0.25 × 0.25 km (with gaps)	1810-2020
M6		✓	✓			✓	0.25 × 0.25 km (with gaps)	1810-2020

Note: This table provides an overview of the different models used in the construction of the spatial crosswalks introduced in this paper. M1 constructs crosswalks based on units' relative areas. M2–M6 are based on sub-county population estimates by Fang and Jawitz (2018) and Leyk and Uhl (2018). M2 uses historical information on urban centers around which it extrapolates population distributions according to a power distribution. M3 additionally excludes non-inhabitable areas, such as swamps, bodies of water, or legally protected areas (e.g., national and state parks) in the population weights. M4 further accounts for the mean elevation when constructing the population-based weights. M5 bases population distributions on historical property records using binary indicators per grid cell for any property built-up in a given period. M6 follows the same approach as M5 but uses the property counts as opposed to mere presence of properties. M5 and M6 only begin in 1810 due to the lack of available property records before that time, with increasing gaps in the available data going further back in time.

reduce the areal coverage of the LU-based crosswalks. To the extent that one cannot construct weights if there are no property records within a given origin county, this means a larger share of origin counties cannot be harmonized for earlier sample decades.[14]

Despite the caveats of each of these approaches to population-based harmonization, it is reassuring that all of the population-based approaches outperform the area-based approach above in Figure 3, while being quite similar to each other in terms of accuracy. At the same time, because of the advantages and limitations of each approach (summarized in Table 3), insofar as population-based crosswalks are appropriate given the factors of study, we recommend using M2–M4 for boundary harmonization involving very early census periods and M6 for more recent ones. Furthermore, we ultimately recommend reporting estimates based on all six weighting models, particularly for earlier periods of study—with the full range of estimates across these models being considered conditional upon the contextual particulars, such as the place and factors of study. While any one weighting model has drawbacks on its own, together they can provide a better understanding of the true estimate in settings where harmonization is required, insofar as economic activity is unevenly distributed in space.

4.2. Concluding remarks

A common problem for spatial researchers involves associating aggregate data from one set of boundaries to another, such as across county boundaries at different points in time or across different contemporaneous units. Existing approaches often use the relative area of overlap between different units to generate and apply weights to stock data for origin units, for the purposes of disaggregating and re-aggregating them to some reference unit. These approaches generally assume a uniform distribution of factors within origin units. In this paper, we develop an alternative approach based on models of historical population distribution by Fang and Jawitz (2018) and Leyk and Uhl (2018), with weights based instead on relative *population* size. This mitigates issues present when economic activity is unevenly distributed within counties.

We use these methods to produce a new set of crosswalks, which relax the uniformity assumption and assign greater weight to areas with greater relative population size within counties. We construct area- and population-based crosswalks for 1790 through 2020, mapping aggregate county-level data across U.S. censuses as well as from counties to congressional districts, whose boundaries are correlated with urban density. We crosscheck our weights using official census data for districts, as applied to the balance tests in Lee, Moretti, and Butler (2004). While all crosswalk-based data replicate their results, data constructed using population-based weights consistently outperform area-based ones in terms of similarity to official data. We hope these methods and crosswalks will be of value to spatial researchers across the social sciences, for whom novel historical data often come pre-aggregated.

Notes

[1] Since 2000, Google Scholar registered more than a quarter million articles involving the term "county level."

[2] To provide a concrete example, take the number of manufacturing firms F in 1880 and compute $F_{C_1^{80}} + F_{C_2^{80}} \times a$ to harmonize this variable to the 1870 boundary for county C^{70}.

[3] Area-based crosswalks cover all admitted U.S. states, while population-based crosswalks are limited to the conterminous U.S., excluding Alaska and Hawaii.

[4] For example, under the first approach, counties from the 1800 U.S. Census are harmonized to CDs for the 4th through 8th Congresses, spanning 1795 through 1804; under the second approach, counties from the 1800 U.S. Census are harmonized to CDs for the 7th through 11th Congresses, spanning 1801 through 1810; and under the third approach, counties from the 1800 U.S. Census are harmonized to CDs for the 8th through 12th Congresses, spanning 1803 through 1812.

[5] Given our setting, we use a "USA Contiguous Albers Equal Area Conic" projection for this.

[6] Note that uniformity does not, however, mean that population need be uniformly distributed *across* counties.

[7] Note that this still assumes some uniformity, *within* urban and rural areas; this is further relaxed in M3 and M4.

[8] Two exceptions for M2–M4 are 1960, for which Fang and Jawitz (2018) lacked urban population data, and 2020, for which no granular population data were available. For 1960, we construct a 1 × 1 kilometer grid cell population distribution map based on census tract population data, from which alternative population-based weights are derived. Appendix Section 2.2 provides more details on the construction of the 1960 population grid. For 2020, we use 2010 population distribution to construct population-based weights. Three exceptions for M5–M6 are 1790, 1800, and 2020. We exclude these models for the former two years and use 2010 settlements and properties to construct these models for 2020.

[9] In the Online Appendix, we describe the process and data used to generate these weights in ArcMap for a given census and Congress year pair.

[10] Of course, any disaggregation into smaller units can introduce error in harmonized data, regardless of the weights used.

[11] Our efforts to reconstruct a high school graduation measure are met with mixed results and differ significantly from the measure in Lee, Moretti, and Butler (2004). We therefore exclude this comparison.

[12] It is worth noting that, since the publication of Lee, Moretti, and Butler (2004), standard practice in applied RD research involves the use of narrow "optimal" bandwidths with linear or quadratic vote share polynomials. Hence, estimates in columns 4 and 5 should be preferred in this application.

[13] In contrast, the binary coding used for M5 may safeguard somewhat against this.

[14] About 5% of weights are undefined for counties in 2010 versus about 25% in 1810 (and, of course, M5 and M6 are not available for 1790 or 1800 at all). One option in cases with missing weights is to define missing weights as zeroes. This would effectively give zero weight to data for all origin counties with too few individuals to have property records.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Autor, D., D. Dorn, G. Hanson, and K. Majlesi. 2020. Importing political polarization? The electoral consequences of rising trade exposure. *American Economic Review* 110 (10):3139–83. doi: 10.1257/aer.20170011.

Bazzi, S., A. Ferrara, M. Fiszbein, T. Pearson, and P. A. Testa. 2023. The other great migration: Southern whites and the new right. *The Quarterly Journal of Economics* 138 (3):1577–647. doi: 10.1093/qje/qjad014.

Bazzi, S., M. Fiszbein, and M. Gebresilasse. 2020. "Frontier culture: The roots and persistence of "rugged individualism" in the United States. *Econometrica* 88 (6):2329–68. doi: 10.3982/ECTA16484.

Beddow, J. M., and P. G. Pardey. 2015. Moving matters: The effect of location on crop production. *The Journal of Economic History* 75 (1):219–49. doi: 10.1017/S002205071500008X.

Berkes, E., E. Karger, and P. Nencka. 2023. The census place project: A method for geolocating unstructured place names. *Explorations in Economic History* 87:101477. doi: 10.1016/j.eeh.2022.101477.

Calderon, A., V. Fouka, and M. Tabellini. 2023. Racial diversity and racial policy preferences: The great migration and civil rights. *The Review of Economic Studies* 90 (1):165–200. doi: 10.1093/restud/rdac026.

Chen, Y. 2015. The distance-decay function of geographical gravity model: Power law or exponential law? *Chaos, Solutions & Fractals* 77:174–89. doi: 10.1016/j.chaos.2015.05.022.

Davis, D. R., and D. E. Weinstein. 2002. Bones, bombs, and break points: The geography of economic activity. *American Economic Review* 92 (5):1269–89. doi: 10.1257/000282802762024502.

Eckert, F., A. Gvirtz, J. Liang, and M. Peters. 2020. "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790," NBER Working Paper No. 26770.

Eeckhout, J. 2004. Gibrat's Law for (All) Cities. *American Economic Review* 94 (5):1429–51. doi: 10.1257/0002828043052303.

Fang, Y., and J. W. Jawitz. 2018. High-resolution reconstruction of the United States human population distribution, 1790 to 2010. *Scientific Data* 5 (1):180067. doi: 10.1038/sdata.2018.67.

Ferrara, A., and P. A. Testa. 2023. Churches as social insurance: Oil risk and religion in the U.S. south. *The Journal of Economic History* 83 (3):786–832. doi: 10.1017/S0022050723000268.

Goodchild, M. F., N. Siu, and Ngan Lam. 1980. Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing* 1 (3):297–312.

Gregory, I. N. 2002. The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems* 26 (4):293–314. doi: 10.1016/S0198-9715(01)00013-8.

Haines, M. 2010. Historical, demographic, economic, and social data: The United States, 1790-2002. Inter-university Consortium for Political and Social Research [distributor], Ann Arbor, MI, 05-21. doi: 10.3886/ICPSR02896.v3., 2010.

Han, Z., H. V. Milner, and K. J. Mitchener. 2023. "The Deep Roots of American Populism," *SSRN Working Paper No. 4523224*.

Hanlon, W. W., and S. Heblich. 2022. History and urban economics. *Regional Science and Urban Economics* 94:103751. doi: 10.1016/j.regsciurbeco.2021.103751.

Hornbeck, R. 2010. Barbed wire: Property rights and agricultural development. *Quarterly Journal of Economics* 125 (2):767–810. doi: 10.1162/qjec.2010.125.2.767.

Hornbeck, R., and S. Naidu. 2014. When the levee breaks: Black migration and economic development in the American south. *American Economic Review* 104 (3):963–90. doi: 10.1257/aer.104.3.963.

Lee, D., S. E. Moretti, and M. J. Butler. 2004. Do voters affect or elect policies? Evidence from the U. S. house. *The Quarterly Journal of Economics* 119 (3):807–59. doi: 10.1162/0033553041502153.

Lee, S., and J. Lin. 2018. Natural amenities, neighborhood dynamics, and persistence in the spatial distribution of income. *The Review of Economic Studies* 85 (1):663–94. doi: 10.1093/restud/rdx018.

Lewis, J. B., B. DeVine, L. Pritcher, and K. C. Martis. 2021. *United States congressional district shapefiles*. Accessed June 30, 2021. https://cdmaps.polisci.ucla.edu/.

Leyk, S., and J. H. Uhl. 2018. HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years. *Scientific Data* 5 (1):180175. doi: 10.1038/sdata.2018.175.

Logan, J. R., B. D. Stults, and Z. Xu. 2016. Validating population estimates for harmonized Census Tract Data, 2000–2010. *Annals of the American Association of Geographers* 106 (5):1013–29. doi: 10.1080/24694452.2016.1187060.

Manson, S., J. Schroeder, D. Van Riper, T. Kugler, and S. Ruggles. 2020. *IPUMS national historical geographic information system," Version 15.0 [dataset]*. Minneapolis, MN. doi: 10.18128/D050.V15.0.

Markoff, J., and G. Shapiro. 1973. The linkage of data describing overlapping geographical units. *Historical Methods Newsletter* 7 (1):34–46. doi: 10.1080/00182494.1973.10112670.

Miguel, E., and G. Roland. 2011. The long-run impact of bombing Vietnam. *Journal of Development Economics* 96 (1):1–15. doi: 10.1016/j.jdeveco.2010.07.004.

Schroeder, J. P. 2016. *Historical Population Estimates for 2010 U.S. States, Counties and Metro/Micro Areas, 1790-2010*. University Digital Conservancy, University of Minnesota Data Repository. doi: 10.13020/D6XW2H.

Testa, P. A. 2021. The economic legacy of expulsion: Lessons from post-war Czechoslovakia. *The Economic Journal* 131 (637):2233–71. doi: 10.1093/ej/ueaa132.