NBER WORKING PAPER SERIES

NEW AREA- AND POPULATION-BASED GEOGRAPHIC CROSSWALKS FOR U.S. COUNTIES AND CONGRESSIONAL DISTRICTS, 1790–2020

Andreas Ferrara
Patrick A. Testa
Liyang Zhou

Working Paper 32206
http://www.nber.org/papers/w32206

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2024

There are no funding or relevant financial relationships to disclose as none were involved in this research. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

New Area- and Population-based Geographic Crosswalks for U.S. Counties and Congressional Districts, 1790–2020
Andreas Ferrara, Patrick A. Testa, and Liyang Zhou
NBER Working Paper No. 32206
March 2024
JEL No. N01,N9,R12

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.

Andreas Ferrara
Department of Economics
University of Pittsburgh
4906 Wesley W. Posvar Hall
Pittsburgh, PA 15260
and NBER
a.ferrara@pitt.edu

Patrick A. Testa
Department of Economics
6823 St Charles Ave
Tilton Hall
New Orleans, LA 70118
United States
ptesta@tulane.edu

Liyang Zhou
Department of Economics
University of Pittsburgh
4514 Wesley W. Posvar Hall
Pittsburgh, PA 15260
liz113@pitt.edu

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, Dorn, Hanson and Majlesi, 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 these 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 such boundary harmonization involves areal interpolation (Markoff and Shapiro, 1973; Goodchild and Lam, 1980; Hornbeck, 2010). We illustrate this approach using an example. Suppose a researcher wishes to harmonize the boundaries of U.S. counties as of 1880 to those from 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, Gvirtz, Liang and Peters, 2020; Testa, 2021; Bazzi, Ferrara, Fiszbein, Pearson and Testa, 2023).

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

[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}.

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 et al. (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 et al. (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 the Geographic Crosswalks

In this section, we describe the methods used to generate our area- and population-based crosswalks, with the intention of providing applied researchers with prerequisite background knowledge and intuition for using the crosswalks. We focus on the construction of the 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 our crosswalks, we use data for county boundaries provided by Manson, Schroeder, Van Riper, Kugler and Ruggles (2020) and CD boundaries from Lewis, DeVine, Pritcher and Martis (2021).

We construct 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 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 Constructing Area-based Crosswalks

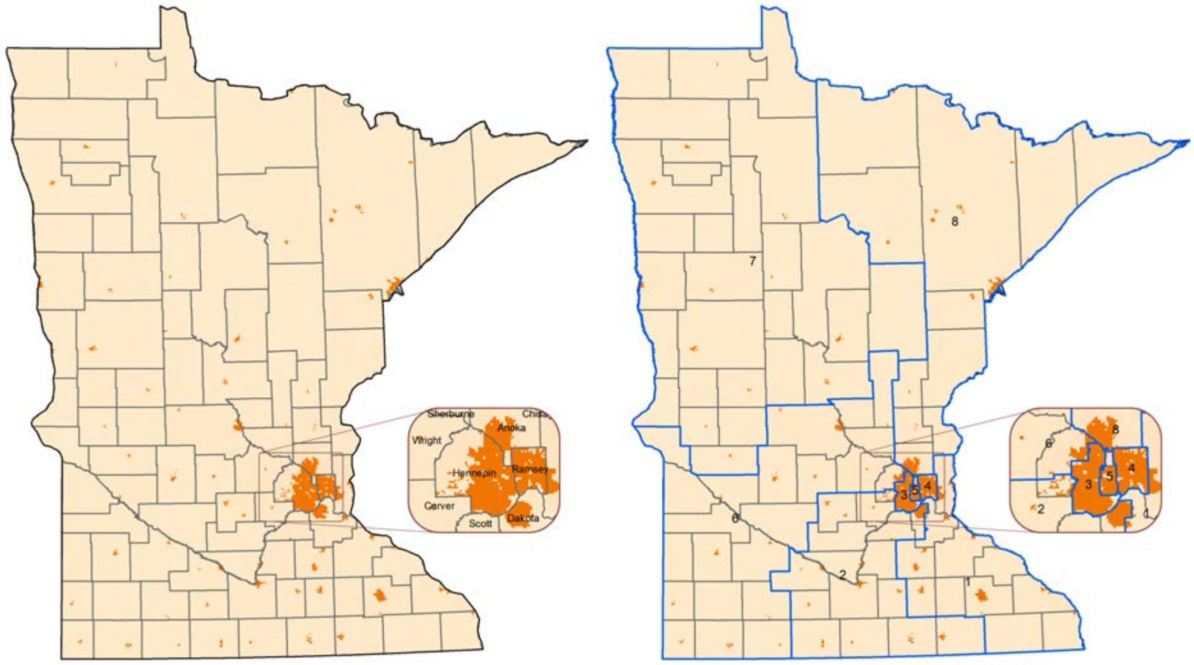
Area-based harmonization procedures generally entail a process of spatial disaggregation and re-aggregation. For our county-to-CD crosswalks, this process involves intersecting a county map from a particular census year with a CD map from a given Congress year. Counties are then disaggregated into a 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 are

[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.

Figure 1: Minnesota Counties, CDs, and Population Distribution Based on 1970 U.S. Census



(a) Counties only

(b) County and CD intersection

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).

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) are 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 weights add up to 1 for each county whose boundaries are being harmonized. A given CD's data values are in turn the aggregates of these weighted values, summed across all counties that have a county-part located in 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.

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 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 1000, 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.

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 $P = (p_1, p_2, ..., p_n)$. Let A be any continuous, two-dimensional subset of C with area $ac \in (0, c)$ and a vector of positive and finite values $R = (r_1, r_2, ..., r_n)$. 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 subpopulation 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.

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

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.

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 United States 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

[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.

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.

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 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.

Constructing the Crosswalks. In order to relax the uniformity assumption, our population-based crosswalks no longer base the disaggregation of county-level data on relative area but rather on relative *population*, using these 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 multiply a county's relevant stock data prior to its aggregation to the CD level.[9]

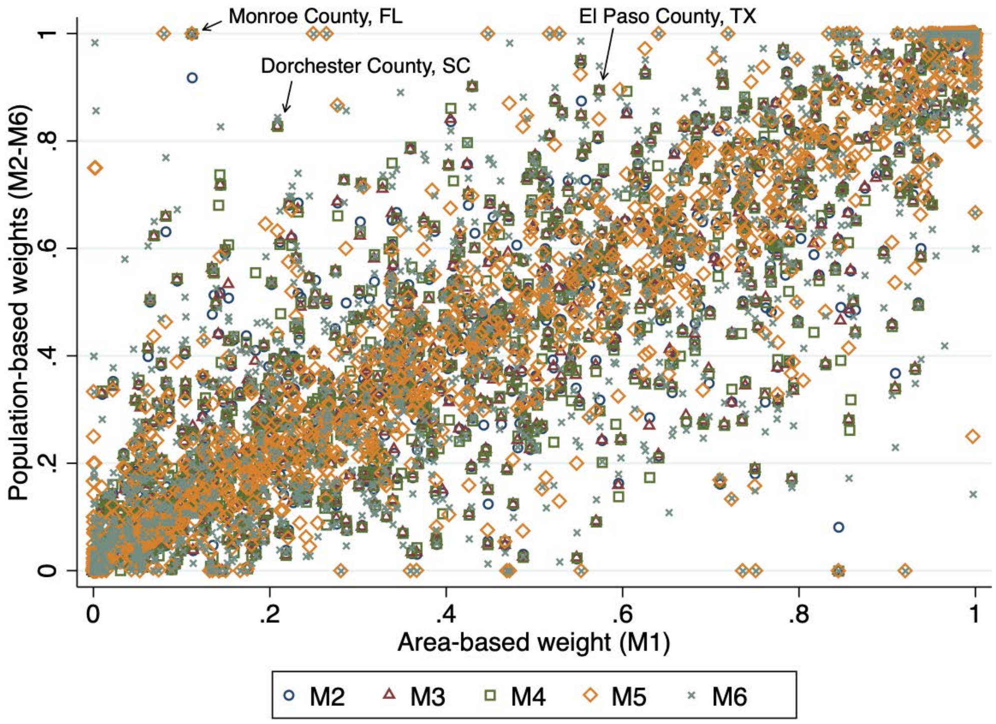
In contrast to the area-based crosswalk, relatively small county-parts in terms of area might in some cases receive a relatively *large* weight—for instance if they are associated with an urban area. Such discrepancies between area- and population-based weights are shown in Figure 2, which relates weights 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% 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.

[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.

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.

2.3 Implementing the 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

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

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.

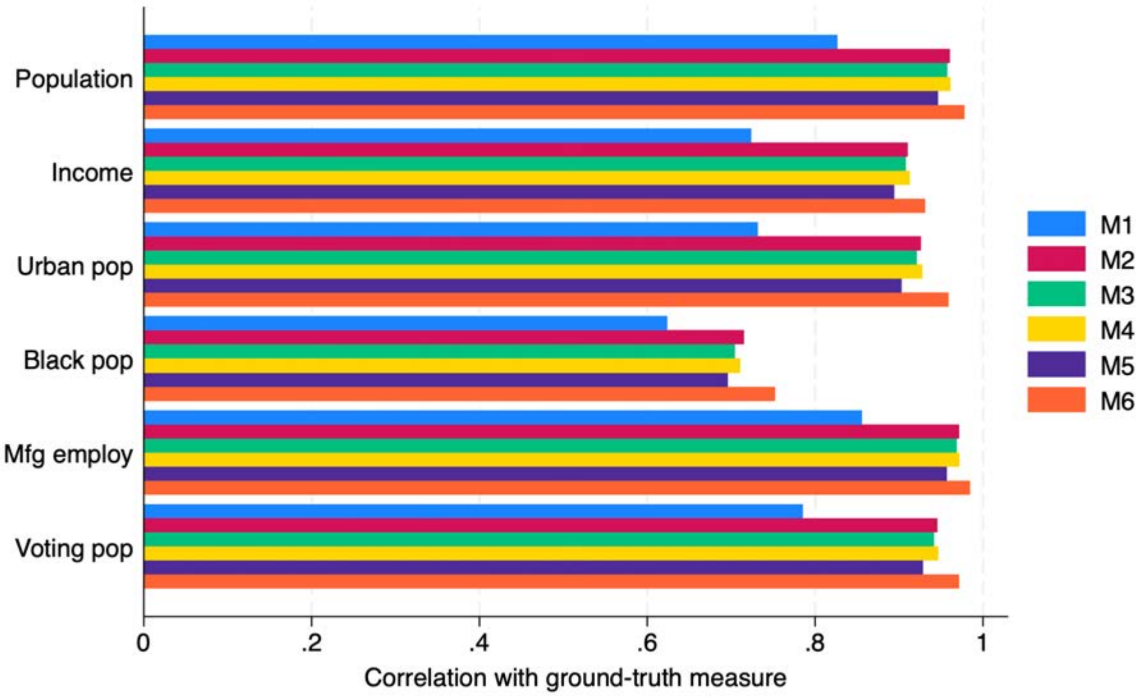
3 Application

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 et al. (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 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 et al. (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 et al. (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 et al. (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%

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

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 et al. (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 et al. (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.

relative to an area-based one. This result mirrors the evaluation of areal interpolation for 2000-10 census tract data in Logan et al. (2016), who show that areal interpolation 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 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 number of Blacks is the most different. This makes sense if you consider where Blacks tend to live in the U.S. In regions like the Midwest and Northeast, such as in states like Illinois, Michigan, and Maryland, Blacks tend to live disproportionately in highly urban areas, relative to the overall population. As a result, both area- and overall population-based crosswalks will tend to underestimate the number of Blacks living in highly urban CDs, allocating some of those counts instead to adjacent CDs. Thus, it is important to keep in mind when harmonizing

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

Model	Population			Income			Urban population		
	(σ = 379,399.9)			(σ = 10,642.2)			(σ = 311,054.5)		
	Corr.	RMSE	RMAE	Corr.	RMSE	RMAE	Corr.	RMSE	RMAE
M1	0.827	259,040.1	347.7	0.724	10,509.2	70.4	0.732	266,936.9	365.6
M2	0.961	110,015.7	235.3	0.910	5,019.8	52.6	0.926	122,646.4	266.7
M3	0.957	114,766.7	240.2	0.908	5,112.1	53.2	0.921	127,017.5	270.8
M4	0.961	109,053.6	231.9	0.913	4,937.3	52.2	0.928	121,398.3	263.5
M5	0.946	133,773.2	274.3	0.894	5,637.2	56.5	0.903	145,714.3	298.1
M6	0.978	83,491.9	186.2	0.931	4,399.6	47.5	0.959	93,615.3	223.1

Model	Black population			Manufacturing employment			Voting population		
	(σ = 83,206.6)			(σ = 46,539.5)			(σ = 229,314.4)		
	Corr.	RMSE	RMAE	Corr.	RMSE	RMAE	Corr.	RMSE	RMAE
M1	0.624	68,357.9	175.1	0.856	28,001.3	110.8	0.785	180,758.4	294.9
M2	0.715	58,798.8	161.9	0.972	11,410.5	78.0	0.946	79,386.8	205.2
M3	0.704	59,888.9	163.2	0.969	11,992.5	79.4	0.942	82,445.3	209.1
M4	0.711	59,300.7	161.6	0.972	11,340.0	77.0	0.947	78,449.2	202.6
M5	0.696	62,006.3	171.9	0.957	14,557.7	89.1	0.929	94,659.7	234.9
M6	0.752	56,392.5	161.4	0.985	8,585.6	68.4	0.971	58,353.2	164.6

Note: This table compares harmonized CD-level data generated by our six crosswalk weights to official CD-level census data, as featured in Lee et al. (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 root 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.

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 et al. (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 replicate the balance tests from Table 2 in Lee et al. (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 replicate the balance test for other variables considered by Lee et al. (2004); as in their paper, % urban and % Black show slight but statistically significant discontinuities across multiple specifications. We report these estimates in Tables A1–A5 in the Online Appendix. Overall, most observable characteristics show

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 et al. (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 percent mark. Column (6) includes a fourth order polynomial in Democratic vote share, which is interacted with the above-below 50 percent dummy. The unit of observation is the district-congress. Standard errors are clustered by district-decade. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

few differences between Democratic and Republican CDs around the tied-election threshold. As such, our estimation, based on data crosswalked to the CD-level from county-level data, replicates the original balance tests in Lee et al. (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 areal interpolation. 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.

[12] It is worth noting that, since the publication of Lee et al. (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.

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.

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).

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 equation (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 non-existence 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 et al. (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 20th 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.

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 areal interpolation. 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.

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 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 appro-

[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.

priate 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 reaggregating 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 et al. (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.

References

Autor, David, David Dorn, Gordon Hanson, and Kaveh Majlesi, "Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure," *American Economic Review*, 2020, *110* (10), 3139–83.

Bazzi, Samuel, Andreas Ferrara, Martin Fiszbein, Thomas Pearson, and Patrick A. Testa, "The Other Great Migration: Southern Whites and the New Right," *Quarterly Journal of Economics*, 2023, *138* (3), 1577–1647.

_ , Martin Fiszbein, and Mesay Gebresilasse, "Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States," *Econometrica*, 2020, *88* (6), 2329–2368.

Beddow, Jason M. and Philip G. Pardey, "Moving Matters: The Effect of Location on Crop Production," *Journal of Economic History*, 2015, *75* (1), 219–49.

Berkes, Enrico, Ezra Karger, and Peter Nencka, "The census place project: A method for geolocating unstructured place names," *Explorations in Economic History*, 2023, *87* (101477).

Calderon, Alvaro, Vasiliki Fouka, and Marco Tabellini, "Racial Diversity and Racial Policy Preferences: The Great Migration and Civil Rights," *Review of Economic Studies*, 2023, *90* (1), 165–200.

Chen, Yanguang, "The distance-decay function of geographical gravity model: Power law or exponential law?," *Chaos, Solutions & Fractals*, 2015, *77*, 174–189.

Davis, Donald R. and David E. Weinstein, "Bones, Bombs, and Break Points: The Geography of Economic Activity," *American Economic Review*, 2002, *92* (5), 1269–1289.

Eckert, Fabian, Andres Gvirtz, Jack Liang, and Michael Peters, "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790," *NBER Working Paper No. 26770*, 2020.

Eeckhout, Jan, "Gibrat's Law for (All) Cities," *American Economic Review*, 2004, *94* (5), 1429–1451.

Fang, Yu and James W. Jawitz, "High-resolution reconstruction of the United States human population distribution, 1790 to 2010," *Scientific Data*, 2018, *5*, https://doi.org/10.1038/sdata.2018.67.

Ferrara, Andreas and Patrick A. Testa, "Churches as Social Insurance: Oil Risk and Religion in the U.S. South," *Journal of Economic History*, 2023, *83* (3), 786–832.

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

Gregory, Ian N., "The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons," *Computers, Environment and Urban Systems*, 2002, *26* (4), 293–314.

Haines, Michael, "Historical, Demographic, Economic, and Social Data: The United States, 1790-2002," *Inter-university Consortium for Political and Social Research [distributor], Ann Arbor, MI, 2010-05-21. https://doi.org/10.3886/ICPSR02896.v3*, 2010.

Han, Ze, Helen V. Milner, and Kris J. Mitchener, "The Deep Roots of American Populism," *SSRN Working Paper No. 4523224*, 2023.

Hanlon, Walker W. and Stephan Heblich, "History and urban economics," *Regional Science and Urban Economics*, 2022, *94* (103751).

Hornbeck, Richard, "Barbed Wire: Property Rights and Agricultural Development," *Quarterly Journal of Economics*, 2010, *125* (2), 767–810.

_ and Suresh Naidu, "When the Levee Breaks: Black Migration and Economic Development in the American South," *American Economic Review*, 2014, *104* (3), 963–90.

Lee, David S., Enrico Moretti, and Matthew J. Butler, "Do Voters Affect or Elect Policies? Evidence from the U. S. House," *Quarterly Journal of Economics*, 2004, *119* (3), 807–859.

Lee, Sanghoon and Jeffrey Lin, "Natural Amenities, Neighborhood Dynamics, and Persistence in the Spatial Distribution of Income," *Review of Economic Studies*, 2018, *85* (1), 663–694.

Lewis, Jeffrey B., Brandon DeVine, Lincoln Pritcher, and Kenneth C. Martis, *United States Congressional District Shapefiles*, 2021, *https://cdmaps.polisci.ucla.edu/ (Accessed on June 30, 2021).*

Leyk, Stefan and Johannes H. Uhl, "HISDAC-US, historical settlement data compilation for the conterminous United States over 200 years," *Scientific Data*, 2018, *5* (180175), 1–14.

Logan, John R., Brian D. Stults, and Zengwang Xu, "Validating Population Estimates for Harmonized Census Tract Data, 2000–2010," *Annals of the American Association of Geographers*, 2016, *106* (5), 1013–1029.

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles, "IPUMS National Historical Geographic Information System," *Version 15.0 [dataset]. Minneapolis, MN. DOI: http://doi.org/10.18128/D050.V15.0.*, 2020.

Markoff, John and Gilbert Shapiro, "The Linkage of Data Describing Overlapping Geographical Units," *Historical Methods Newsletter*, 1973, *7* (1), 34–46.

Miguel, Edward and Gerard Roland, "The long-run impact of bombing Vietnam," *Journal of Development Economics*, 2011, *96* (1), 1–15.

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

Testa, Patrick A., "The Economic Legacy of Expulsion: Lessons from Post-War Czechoslovakia," *Economic Journal*, 2021, *131* (637), 2233–2271.

Online Appendix for
"New Area- and Population-based Geographic Crosswalks for U.S. Counties and Congressional Districts, 1790–2020"[*]

Andreas Ferrara[†] Patrick A. Testa[‡] Liyang Zhou[§]

February 27, 2024

Table of Contents

[*] The area- and population-based crosswalks produced in the paper, teaching material, as well as code and data for the replication exercise can be downloaded at https://doi.org/10.3886/E150101.

[†] University of Pittsburgh, Department of Economics and NBER. Email: a.ferrara@pitt.edu.

[‡] Tulane University, Department of Economics. Email: ptesta@tulane.edu.

[§] University of Pittsburgh, Department of Economics. Email: liyang.zhou@pitt.edu.

1 Overview of Data

1.1 Data Used to Produce Crosswalks

Our crosswalks are based in part on the U.S. county shapefiles from https://www.nhgis.org/. For the period 1790 to 2000, we use the shapefiles based on the 2000 TIGER/Line county boundary definitions. For 2010 and 2020, these are not available. We therefore use contemporaneous TIGER/Line shapefiles for those years instead. The shapefiles for 2010 and 2020 are sourced from https://www.census.gov/.

Shapefiles for congressional districts (CD) are from Lewis, DeVine, Pritcher and Martis (2021) for the 1st to 114th Congress, as available at https://cdmaps.polisci.ucla.edu/. The shapefiles for CDs from the 115th and 116th Congress are obtained from https://catalog.data.gov/dataset/tiger-line-shapefile-2016-nation-u-s-115th-congressional-district-national.

Historical population distribution maps used to construct M2–M4 come from Fang and Jawitz (2018). These are described in greater detail in the next section. We use U.S. Census Bureau tract shapefiles and population data for 1960 to construct a population distribution raster for that year, both from NHGIS, since Fang and Jawitz (2018) do not have urban population data for that year.

Historical property records maps used to construct M5–M6 come are based on Leyk et al. (2020) and available for download at Uhl and Leyk (2020a) and Uhl and Leyk (2020b). These are described in greater detail in Section 3.

1.2 Data Used in Replication Exercise

For the replication of Lee et al. (2004) in the application section of the main paper, we use the data and code to replicate Table 2 from Enrico Moretti's website. The data can be accessed via https://eml.berkeley.edu/~moretti/data3.html. Their replication files include the data from the U.S. Census extracts for 1960-90. To reconstruct CD level from county-level census data, we use data from the U.S. Census of Population and Housing for 1960 and 1970 from Haines (2010), as well as with information from the County and City Data Books for the years 1962, 1967, 1972 1977, 1983, and 1994. This allows us to test the performance of different crosswalks (area- versus population-based) when crosswalking the county-level data to the CD-level, in comparison to the official CD-level data produced by the Census Bureau. Urban population data for 1990 come from the 1990 U.S. Census, which is separately sourced from NHGIS.

2 Overview of Spatial Models in Fang and Jawitz (2018)

2.1 Data

To estimate spatial models of population distribution for the conterminous U.S., Fang and Jawitz (2018) use the following data:

1. Total and urban (threshold 2,500+) population at the county level from the National Historical Geographic Information System (https://www.nhgis.org/) between 1790 and 2010 (excluding 1960, as these data are missing, and 2020, as these are not yet available),

2. Data on water bodies from the National Hydrography Dataset (http://nhd.usgs.gov/) including lakes, ponds, marsh and swamp land, and other minor water bodies as of 2001,

3. Boundaries of protected areas from the National Gap Analysis Program (http://gapanalysis.usgs.gov/padus/) as of 2012,

4. Elevation data from the NASA Shuttle Radar Topography Mission Version 3.0 (http://www2.jpl.nasa.gov/srtm/) as of 2013,

for 7,754,146 square-kilometer grid cells in the conterminous United States. They caution that protected areas were established in more recent times and that areas settled by American Natives will undercount population as this group was only fully included in the U.S. censuses from 1900 onward. An illustration of the spatial variation of their data for the year 2000 is shown in Figure A2.

Defining the Spatial Extent of Urban Areas

Fang and Jawitz (2018) measure the areal extent of 3,610 urban areas using the 2000 U.S. Census and project area backward in time using a power law relationship between an urban area's population and its spatial extent,

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

where $A_{U,\varphi}$ is the spatial extent of urban area and urban areas are indexed by φ in U.S. Census Bureau division δ, α_δ and β_δ are the coefficients of the power function, which are assumed to be constant over time,[1] and $P_{U,\varphi}$ is the population size. Using the log-transformed version of the model and historical population data from the census, they then estimate the historical area size of the urban areas in their sample.

Defining Inhabitable Areas

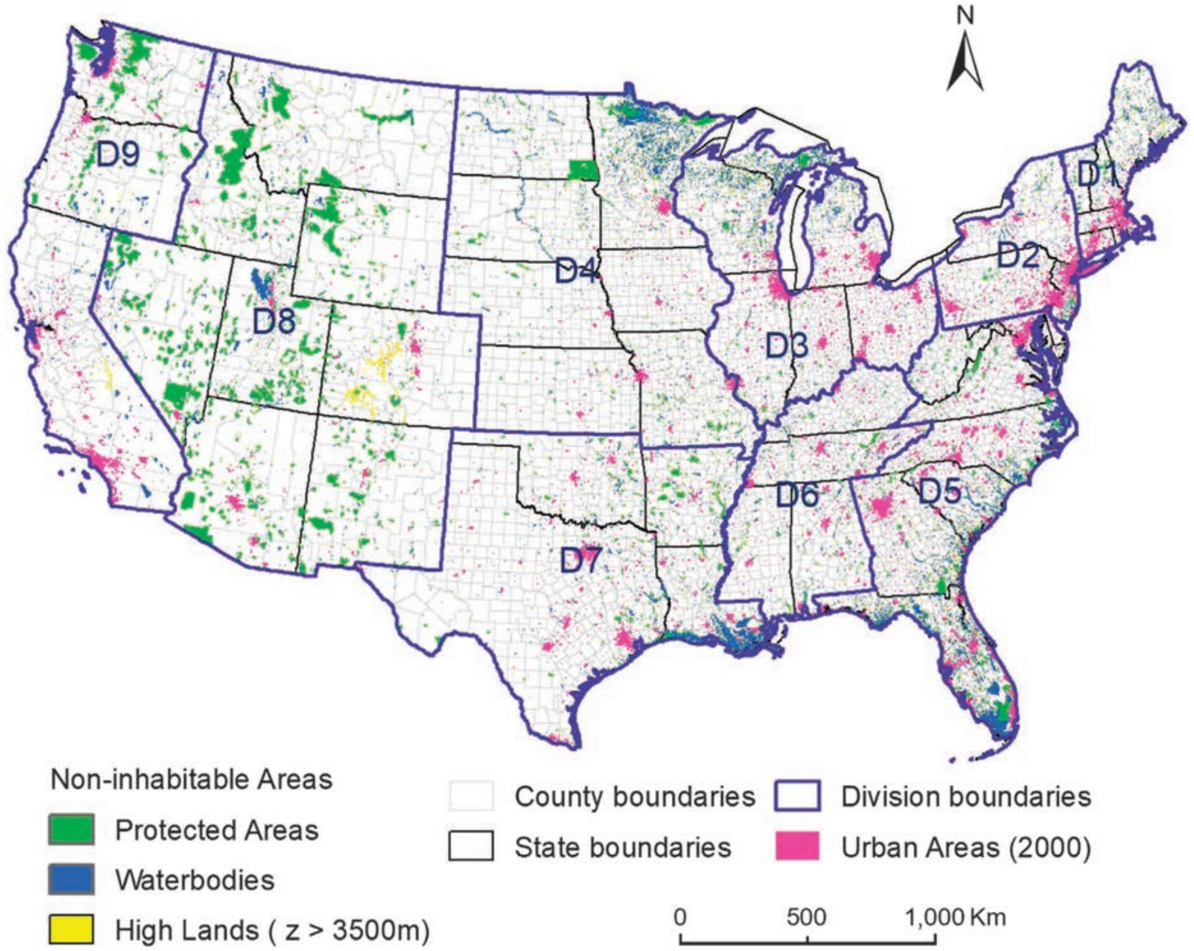
Fang and Jawitz (2018) define inhabitable areas as those that are not water bodies larger than 1 square kilometer, protected areas, or areas with an elevation of more than 3,500 meters.

2.2 Models

Fang and Jawitz (2018) employ five spatial models of population distribution for the conterminous U.S. The first model corresponds to county-level population distributions. This assumes a

[1] They motivate this assumption referencing other historical models of urbanization and urban extent in England, China, Pre-Hispanic Mexico, and Japan.

Figure A1: Geographic Distribution of Data Features Used by Fang and Jawitz (2018) for the Year 2000



Note: Map showing variation in the main data sources used by Fang and Jawitz (2018) for the year 2000. D1-D9 refer to U.S. Census Bureau divisions. Source: Figure 1 in (Fang and Jawitz, 2018, p. 3).

uniform population distribution within counties, with each one-by-one kilometer grid cells having the same population value within a given county. The second model differentiates between rural and urban areas, using the urban areal extents and urban population stock data described above. Within a county, urban population is distributed uniformly within urban areas and the remaining non-urban population is distributed uniformly within rural areas. The third model also does this but first excludes non-inhabitable areas, as described above. The fourth model extends the third by also multiplying the population raster by topographic suitability weights. These four models correspond to our M1–M4, respectively. We do not discuss or utilize their fifth model, which incorporates information from a constructed “socioeconomic desirability” index. For most applications in the social sciences this allocation of population is problematic as it is based on potentially endogenous variables, such as the distance from the periphery to the core of cities.[2]

Alternative Models for 1960

Because digital urban population data are missing for the 1960 U.S. Census, Fang and Jawitz (2018) do not develop models for that year. In lieu of this, and in the interest of completeness, we still construct crosswalks based on the area-based M1 for 1960, assuming a uniform population distribution within counties. We also construct crosswalks based on an alternative M2, in which we use the most granular population data possible: (i) population and boundary data at the 1960 U.S. Census tract level, where available (coverage is most urban areas), and (ii) population and boundary data at the 1960 county level otherwise. This is akin to adopting a uniformity assumption within rural counties, i.e., where population distribution is likely to be relatively homogeneous anyway. These population and boundary data are transformed into a raster with square kilometer grid cells, to match the Fang and Jawitz (2018) M2 for other years. In the crosswalk file, these weights take the places of M2–M4 for 1960.

Modeling Population Distribution for 2020

Given that Fang and Jawitz (2018) predates the 2020 U.S. Census, and urban or census-tract-level population data are not yet available, we utilize urban population data from 2010 and the corresponding models (M2–M4) from Fang and Jawitz (2018) in order to model population distributions within counties for 2020 counties that are being harmonized to CDs or to other counties.

3 Overview of Spatial Models in Leyk et al. (2020)

To construct maps of historical settlements for the conterminous U.S., Leyk et al. (2020) use settlement layers from the Historical Settlement Data Compilation for the United States (HISDAC-US), which is itself derived from historical property records found in the Zillow Transaction and Assessment Database (ZTRAX). These settlement layers are granular, based on 250×250 meter grid cells, and built for 5 year periods beginning in 1810. The database includes a variety of measures:

- The historical "built-up areas" measure (BUA) defines for each grid cell the presence of at least one built-up structure in a given year (built-up$= 1$).
- The historical "built-up property records" measure (BUPR) defines for each grid cell the number of built-up property records in a given year.
- The historical "built-up property locations" measure (BUPL) defines for each grid cell the number of unique locations of built-up property records in a given year (i.e., independent of ownership). This is highly correlated with the BUPR measure.

[2] Unlike geographic and topographic features, distance based on a gravity model is also likely to change the most over time due to changes in available transportation methods.

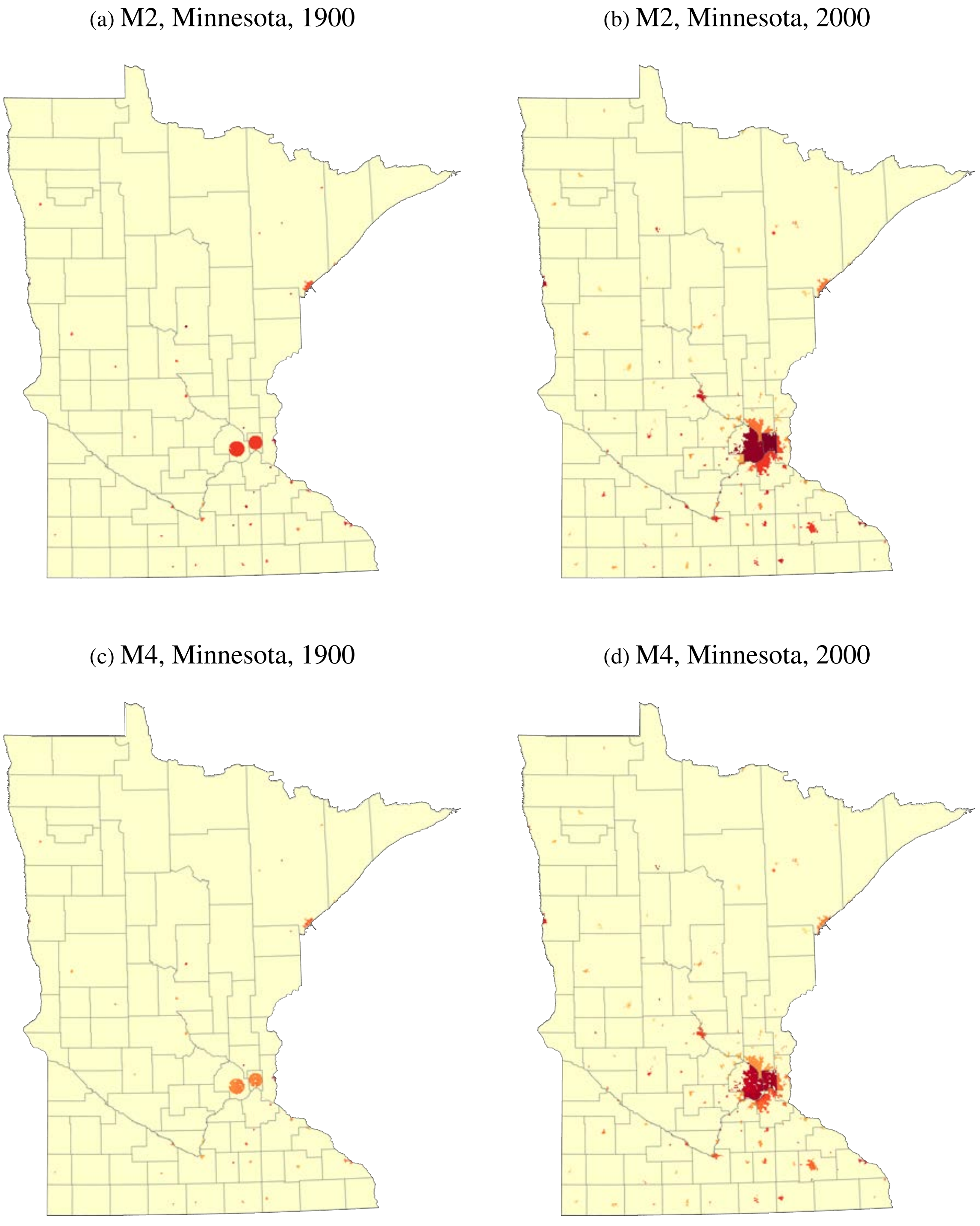
- The “first built-up year” measure (FBUY) defines for each grid cell the earliest construction year on record.

The first two measures are used to construct and thus correspond to our models M5 and M6, respectively, for all census years 1810 through 2020. As with M1–M4, we use the 2010 map to construct our models for 2020.

Built-up property records are importantly highly correlated with population size, as the authors show in Leyk et al. (2020). 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. On that basis, property records provide an accurate and granular proxy for historical population counts.

Figures A2 and A3 provide different visualizations of the weights to construct our crosswalks in our example of Minnesota. Figure A2 shows the population-based weights for the M2 (panels a and b) and M4 (panels c and d) models in 1900 and 2000, respectively. Figure A3 maps the spatial extent of built-up areas (panels a and b) and the built-up property records (panels c and d) for the same state and time periods.

Figure A2: Examples of M2 and M4 from Fang and Jawitz (2018), Used to Construct Our M2 and M4 Weights



(a) M2, Minnesota, 1900

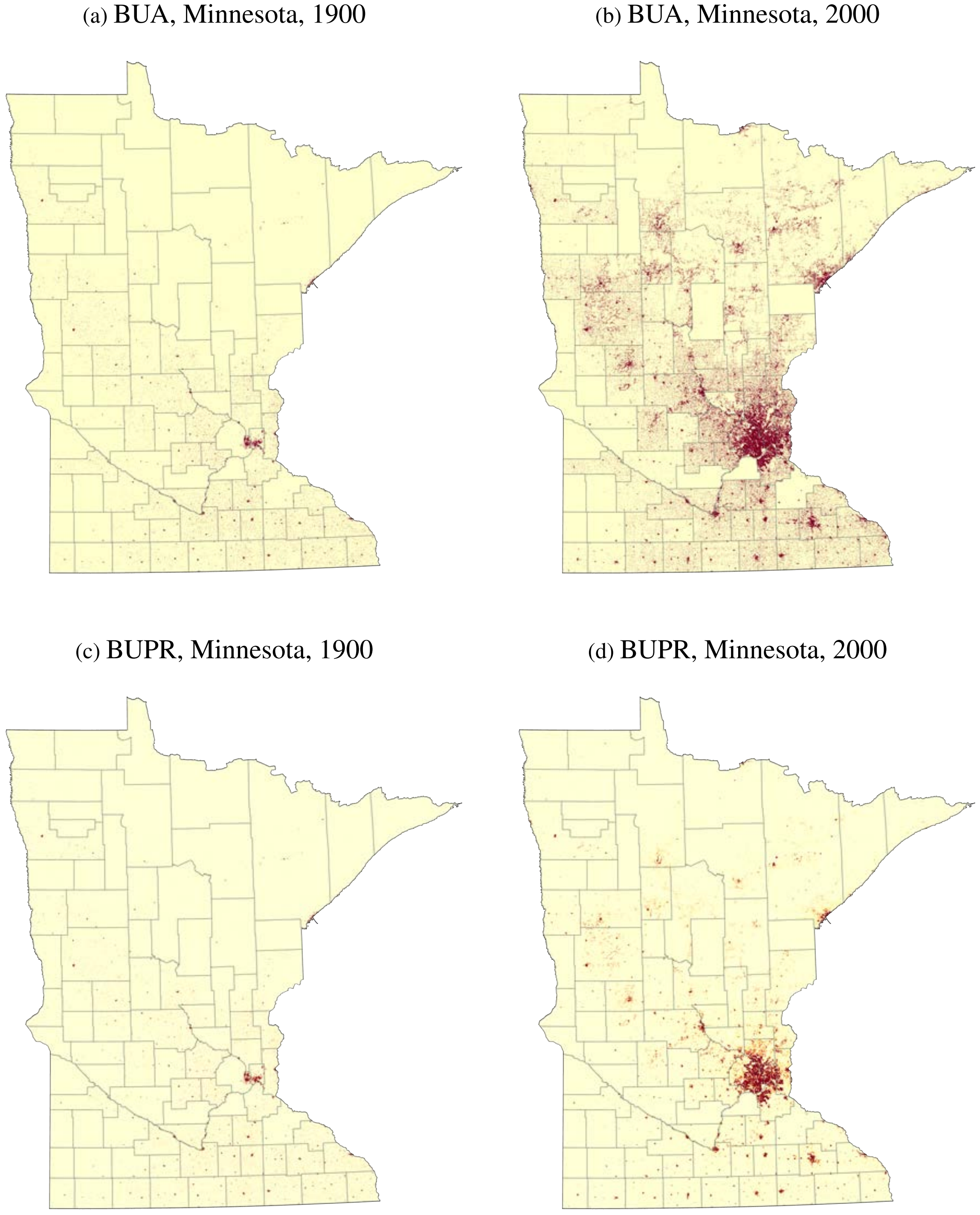
(b) M2, Minnesota, 2000

(c) M4, Minnesota, 1900

(d) M4, Minnesota, 2000

Note: This figure shows the land area of the state of Minnesota with built-up distribution information for 1900 and 2000, where darker orange a greater number of residents per square kilometer for M2 and M4 (used to construct our weights of the same names). The gray boundaries show the state's county boundaries as of the 2010 U.S. Census. County shapefiles are from Manson, Schroeder, Van Riper, Kugler and Ruggles (2020). Population distribution estimates for 1900 and 2000 are from Fang and Jawitz (2018).

Figure A3: Examples of Built-Up Areas (BUA) and Built-Up Property Records (BUPR) from Leyk et al. (2020), Used to Construct Our M5 and M6 Weights



(a) BUA, Minnesota, 1900

(b) BUA, Minnesota, 2000

(c) BUPR, Minnesota, 1900

(d) BUPR, Minnesota, 2000

Note: This figure shows the land area of the state of Minnesota with built-up distribution information for 1900 and 2000, where darker orange implies at least one build-up property within the grid cell by the year for BUA (used to construct our M5), or a greater number of individual built-up property record by the year for BUPR (used to construct our M6). County shapefiles are from Manson et al. (2020). Built-up property distribution information for 1900 and 2000 are from Leyk et al. (2020).

4 Construction of the Geographic Crosswalks

This paper makes a number of contributions to the study of spatial phenomena in the social sciences, with a wide range of applications for urban economists, political scientists, and economic historians, among others. Most notably, we extend existing area-based approaches to harmonizing county boundaries across U.S. censuses over time, by relaxing standard uniformity assumptions regarding population distribution within counties for the contiguous U.S. (i.e., excluding Alaska and Hawaii). To do this, we rely on maps based on historical population estimates for 1×1 kilometer grid cells from Fang and Jawitz (2018) (for M2–M4) and maps based on historical property records for 250×250 meter grid cells from Leyk et al. (2020) (for M5–M6). These capture sub-county variation in the population distribution for counties being harmonized, to the extent that more or less of its population may be located in some "parts" relative to others. Given that harmonization frequently involves spatial disaggregation of counties prior to the re-aggregation to the boundaries of some "reference" unit, this better ensures that initial county stock variables will be properly weighted in the disaggregation process prior to being re-aggregated to different reference county boundaries.

We then apply this innovation in order to construct wholly original county-to-congressional district (CD) crosswalks, spanning the entirety of U.S. congressional history, from 1790 to 2020. These crosswalks can be used to aggregate county-level data to the CD-level. Because CD boundaries often do not align with county boundaries, county data must often be spatially disaggregated before being re-aggregated to the CD level. This renders the models of historical sub-county population distribution from Fang and Jawitz (2018) and Leyk et al. (2020) quite important.

We now describe this process of disaggregation and re-aggregation, used to generate our county-to-county and county-to-CD crosswalks, beginning with the area-based crosswalks used previously in the literature.

4.1 Area-based Crosswalks

Area-based harmonization procedures entail a simple process of spatial disaggregation and re-aggregation. To construct our county-to-CD crosswalks, this involves intersecting a county map from a particular census year with a CD map from a particular Congress year. Counties are then disaggregated into a set of sub-county units ("county-parts"), based the CD in which they are located. We then calculate the areas (in square meters) of all counties, all CDs, and all county-parts, based on a "USA Contiguous Albers Equal Area Conic" projection. Once counties are disaggregated based on CD intersections, county-parts are re-aggregated based on their CD, with the sum of the areas of the county-parts matching the area of the whole CD. This is the same process outlined in Eckert, Gvirtz, Liang and Peters (2020) and used to construct the county-to-county crosswalks featured in Hornbeck (2010) and Bazzi, Fiszbein and Gebresilasse (2020), among others.

How are the various data values of the initial counties (e.g., total population, total number of Blacks) associated with CDs in this process? Under an area-based procedure, each county-part is assigned each of its county's data values, weighted by the share of the county's total *area* that belongs to that county-part. These weights add up to 1 for each county. A given CD's data values are in turn the aggregates of these weighted values, summed across all counties that have a county-part located in that CD. For our area-based weights to be appropriate in settings where county and CD boundaries overlap, the following condition and proposition are relevant:

Proposition 1. *Suppose all counties satisfy:*

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

Then our area-based crosswalk will accurately map county-level values to the congressional district level for all districts.

Proof. Let D be any continuous, two-dimensional congressional district with area $d > 0$ and a vector of positive and finite values $Q = (q_1, q_2, ..., q_n)$. Suppose D can be decomposed into 1 county-part each from M counties $j = 1, ..., m$, each with finite area $a_j c_j$ and a vector of positive and finite values $R_j = (r_{j1}, r_{j2}, ..., r_{jn})$, such that:

$$\sum_{j=1}^{M} a_j c_j = d,$$

$$\sum_{j=1}^{M} R_j = Q,$$

where a_j is the share of county j's area belonging to its county-part that lies in D. As such, $(a_1, a_2, ..., a_m)$ is the vector of weights associated with our area-based crosswalk, which map values from counties $j = 1, ..., m$ to D.

Yet suppose in actuality that our area-based crosswalk does not accurately map county-level values to D, such that:

$$Q \neq \sum_{j=1}^{M} a_j P_j,$$

where P_j is the vector of positive and finite values associated with county j. It follows that $a_j P_j \neq R_j$ for at least one county j, a violation of uniformity. □

4.2 Population-based Crosswalks

We then construct population-based crosswalks, which rely on a relaxation of the uniformity assumption. To do this, we use information on historical sub-county population distribution by Fang and Jawitz (2018) and Leyk et al. (2020). The former provide 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 a power law scaling relationship; (ii) model 3 (M3), which is is based on a version of M2 that first excludes non-inhabitable areas, such as bodies of water; and (iii) model 4 (M4), which is based on a version of M3 that also weights population counts based on topographic suitability. In contrast, Leyk et al. (2020) derive proxies for historical population size for 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 models: (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. We describe these data and models in detail above.

To relax the uniformity assumption, we no longer base the disaggregation of county-level data on relative area but rather relative population size. These various population distribution raster maps let us calculate the total population counts (or property-based proxy) of each county polygon for each census year, as well as approximate counts for each county-part polygon within a county that lies in a different CD. As with our area-based crosswalk, the ratio of these counts in the county-part relative to the entire "origin" county provides a weight with which to multiply a county's stock data prior to its aggregation to the CD level.

4.3 Using ArcMap and Stata to Construct Crosswalks

Although our crosswalks currently cover the entirety of U.S. congressional history and of the U.S. censuses, time will render them incomplete. We are therefore describing the data and steps taken to (i) generate necessary spatial data and (ii) construct crosswalk weights for a given county-CD pair. In Section 5, we will discuss step-by-step how to apply these weights and harmonize county boundaries to those of a contemporaneous CD. For further illustration, we also include relevant samples of Stata and R code in Section 5.

4.3.1 Guide for Generating Spatial GIS Data

We use ArcMap to generate the necessary spatial information for crosswalks, based on the data sources described in the first part of this Online Appendix. This process goes as follows for each county-CD crosswalk:

We first calculate the area (in square meters) of each county polygon within each county shapefile. Working within a NAD 1983 data frame, all area calculations are based on a “USA Contiguous Albers Equal Area Conic” projection. We then intersect the county shapefile with a given CD polygon shapefile, using the “intersect” tool. The output table from this intersection provides the full set of county-parts for each county.

For area-based crosswalks, we calculate the area (in square meters) of each county-part. For population-based crosswalks, we use the “zonal statistics table” tool to calculate the sum of population (or property-based proxy) counts within each county-part, based on the raster grid cells for a given model M2–M6 in a given census year. These produce five separate tables, one each for M2–M6, with the same county-part identifiers used in the intersection output table. We then export all six tables to CSV.

In Stata, we import each CSV separately. We then use the identifiers for the county-parts from ArcMap to merge the M2, M3, M4, M5, M6 and intersection output tables. Total population (or property-based proxy) counts for each county for M2–M6 are calculated by summing said counts of all county-parts within each given county.

To generate weights for area-based crosswalks, we calculate the ratio of county-part-to-origin-county area for each county-part. To generate weights for each population-based crosswalks, population (or property-based proxy) counts are divided by the total counts for the origin county for each population-based model. Miscellaneous minor edits are made to make sure county and CD identifiers are consistent across crosswalk years and to fix a few errors relating to duplicate county-parts (e.g., in the case of pre-1960s CDs in states with both at-large and within-state CDs overlapping each other).

5 Step-by-step Guide for Applying the Crosswalks in Stata and R

Below we provide a simple example using Stata and R to show step-by-step how to crosswalk county level aggregates to congressional district boundaries. Here we wish to crosswalk total population and Black population (in levels) from 1960 to the boundaries of the 88th Congress.

First, take total population and Black population at the county level and prepare the data for merging with the crosswalk file.

```
. use state county level statefip counfip var3 using "1962_cnty_and_city_data_book.dta"
(Historical, Demographic, Economic, and Social Data: The United States, 1790-2002)

.
. gen census = 1960

.
. * merge with county level data on Black population
. merge 1:1 state county using "1960_census_county.dta", keepusing(negmtot negftot)

    Result                           # of obs.
    -----------------------------------------

    not matched                             2
        from master                         2  (_merge==1)
        from using                          0  (_merge==2)

    matched                             3,183  (_merge==3)
    -----------------------------------------
```

```
. drop _merge

.
. * keep counties only (level 1) and drop D.C. (no congressional representation)
. keep if level==1 & state!=98
(53 observations deleted)

.
. * generate the variables of interest in levels
. gen black = negmtot + negftot

. gen totpop = var3

.
. * keep only relevant variables for the cross walk
. keep state county census black totpop

.
. * rename county identifiers for merging with the crosswalk
. ren (state county) (icpsrst icpsrcty)
```

Second, use the county and state identifiers to merge the crosswalk file to the data in a 1:m merge. This expands the set of counties to the full set of county-parts belonging to each congressional district intersecting a given county.

```
. * Merge with crosswalks
. merge 1:m icpsrst icpsrcty using "Crosswalk_1960_88.dta"

    Result                           # of obs.
    -----------------------------------------
    not matched                             2
        from master                         1  (_merge==1)
        from using                          1  (_merge==2)

    matched                             7,368  (_merge==3)
    -----------------------------------------

. drop _merge
```

Third, multiply each stock variable with the relevant weights, using the weights from all six models and keeping track of county-parts with missing data or undefined weights.

```
. * Weight county count data by weights
. qui ds black totpop

. foreach v in `r(varlist)´ {
. * If any one "county part" has a missing value, may want to mark the whole of the
. * aggregated district to have a missing value as well, especially if that county
. * part makes up a sizable part of the district
  1.         replace `v´ = -999999999999999 if(`v´==.) & cnty_part_area/cd_area>0.05
.         *Apply weights (for all 4 models)
  2.         gen `v´_m1 = `v´*m1_weight
  3.         gen `v´_m2 = `v´*m2_weight
  4.         gen `v´_m3 = `v´*m3_weight
  5.         gen `v´_m4 = `v´*m4_weight
  6.         gen `v´_m5 = `v´*m5_weight
  7.         gen `v´_m6 = `v´*m6_weight
. }
(0 real changes made)
(2 missing values generated)
(36 missing values generated)
```

```
(36 missing values generated)
(36 missing values generated)
(642 missing values generated)
(642 missing values generated)
(0 real changes made)
(2 missing values generated)
(36 missing values generated)
(36 missing values generated)
(36 missing values generated)
(642 missing values generated)
(642 missing values generated)

.
. qui ds black* totpop*

. foreach v in `r(varlist)´ {

. * If any one "county part" has a missing value due to a missing weight,
. * may want to mark the whole of the aggregated district to have a missing
. * value as well
.         replace `v´ = -999999999999999 if(`v´==.) & cnty_part_area/cd_area>0.05
. }
(0 real changes made)
(1 real change made)
(12 real changes made)
(12 real changes made)
(12 real changes made)
(167 real changes made)
(167 real changes made)
(0 real changes made)
(1 real change made)
(12 real changes made)
(12 real changes made)
(12 real changes made)
(167 real changes made)
(167 real changes made)
```

Fourth, collapse the weighted data on the CD identifier. The unit of observation is now the congressional district.

```
. * Collapse by congressional district
. collapse (sum) black_m* totpop_m*, by(census congress cd_state cd_statefip ///
>                                             cd_stateicp district id cd_area)

.
. * Correct districts with missing "county parts" and otherwise round to nearest integer
. qui ds black_m* totpop_m*

. foreach v in `r(varlist)´ {
  2.
.         replace `v´ = . if(`v´<=0)
  3.         replace `v´ = round(`v´)
  4.
. }
(1 real change made, 1 to missing)
(387 real changes made)
(3 real changes made, 3 to missing)
(160 real changes made)
(3 real changes made, 3 to missing)
(160 real changes made)
(3 real changes made, 3 to missing)
(160 real changes made)
(3 real changes made, 3 to missing)
(150 real changes made)
```

```
(3 real changes made, 3 to missing)
(150 real changes made)
(1 real change made, 1 to missing)
(383 real changes made)
(3 real changes made, 3 to missing)
(159 real changes made)
(3 real changes made, 3 to missing)
(159 real changes made)
(3 real changes made, 3 to missing)
(159 real changes made)
(3 real changes made, 3 to missing)
(150 real changes made)
(3 real changes made, 3 to missing)
(144 real changes made)
```

The final step also rounds the relevant variables or replaces them as missing in cells where this is appropriate.

The following R code below reproduces the example. The logic should carry over to other statistical packages and programming languages.

```
# load relevant libraries
library(haven)
library(dplyr)

# load the required data sets and crosswalks
census_county_1960 <- read_dta('./Data/LMB/County/1960_census_county
   .dta')
cnty_city_data_book_1962 <- read_dta('./Data/LMB/County/1962_cnty_
   and_city_data_book.dta')
crosswalk <- read_dta('./Data/Final␣Crosswalks/Stata/DOA/Crosswalk_
   1960_88.dta')

# Step 1: set up the data frame, generate relevant variables
# keep only counties (level = 1)
df <- cnty_city_data_book_1962 %>%
mutate(census = 1960) %>%
full_join(census_county_1960) %>%
filter(level == 1 & state != 98) %>%
mutate(black = negmtot + negftot, totpop = var3) %>%
select(state, county, census, black, totpop) %>%
rename(icpsrst = state, icpsrcty = county)

# Step 2: merge the crosswalk file to the data
df <- df %>%
full_join(crosswalk, relationship = "one-to-many")
```

```
# Step 3: multiply stock variables with the relevant weights
# do this using the weights from the six different models
df <- df %>%

mutate(black = replace(black, is.na(black) & cnty_part_area/cd_area
   > 0.05, -999999999999999),
totpop = replace(totpop, is.na(totpop) & cnty_part_area/cd_area >
   0.05, -999999999999999)) %>%

mutate(black_m1 = black*m1_weight,
black_m2 = black*m2_weight,
black_m3 = black*m3_weight,
black_m4 = black*m4_weight,
black_m5 = black*m5_weight,
black_m6 = black*m6_weight,
totpop_m1 = totpop*m1_weight,
totpop_m2 = totpop*m2_weight,
totpop_m3 = totpop*m3_weight,
totpop_m4 = totpop*m4_weight,
totpop_m5 = totpop*m5_weight,
totpop_m6 = totpop*m6_weight) %>%

mutate(black_m1 = replace(black_m1, is.na(black_m1) & cnty_part_area
   /cd_area > 0.05, -999999999999999),
black_m2 = replace(black_m2, is.na(black_m2) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
black_m3 = replace(black_m3, is.na(black_m3) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
black_m4 = replace(black_m4, is.na(black_m4) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
black_m5 = replace(black_m5, is.na(black_m5) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
black_m6 = replace(black_m6, is.na(black_m6) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
totpop_m1 = replace(totpop_m1, is.na(totpop_m1) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
totpop_m2 = replace(totpop_m2, is.na(totpop_m2) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
totpop_m3 = replace(totpop_m3, is.na(totpop_m3) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
totpop_m4 = replace(totpop_m4, is.na(totpop_m4) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
totpop_m5 = replace(totpop_m5, is.na(totpop_m5) & cnty_part_area/cd_
   area > 0.05, -999999999999999),
totpop_m6 = replace(totpop_m6, is.na(totpop_m6) & cnty_part_area/cd_
   area > 0.05, -999999999999999))

# Step 4: collapse the weighted data on the CD identifier.
# The unit of observation is now the congressional district.
df <- df %>%

group_by(census, congress, cd_state, cd_statefip, cd_stateicp,
   district, id, cd_area) %>%

summarise(black_m1 = sum(black_m1, na.rm = T),
black_m2 = sum(black_m2, na.rm = T),
black_m3 = sum(black_m3, na.rm = T),
black_m4 = sum(black_m4, na.rm = T),
black_m5 = sum(black_m5, na.rm = T),
black_m6 = sum(black_m6, na.rm = T),
totpop_m1 = sum(totpop_m1, na.rm = T),
totpop_m2 = sum(totpop_m2, na.rm = T),
totpop_m3 = sum(totpop_m3, na.rm = T),
totpop_m4 = sum(totpop_m4, na.rm = T),
totpop_m5 = sum(totpop_m5, na.rm = T),
totpop_m6 = sum(totpop_m6, na.rm = T)) %>%

ungroup() %>%

mutate(black_m1 = ifelse(black_m1 <= 0, NA, round(black_m1)),
black_m2 = ifelse(black_m2 <= 0, NA, round(black_m2)),
black_m3 = ifelse(black_m3 <= 0, NA, round(black_m3)),
black_m4 = ifelse(black_m4 <= 0, NA, round(black_m4)),
black_m5 = ifelse(black_m5 <= 0, NA, round(black_m5)),
black_m6 = ifelse(black_m6 <= 0, NA, round(black_m6)),
totpop_m1 = ifelse(totpop_m1 <= 0, NA, round(totpop_m1)),
totpop_m2 = ifelse(totpop_m2 <= 0, NA, round(totpop_m2)),
totpop_m3 = ifelse(totpop_m3 <= 0, NA, round(totpop_m3)),
totpop_m4 = ifelse(totpop_m4 <= 0, NA, round(totpop_m4)),
totpop_m5 = ifelse(totpop_m5 <= 0, NA, round(totpop_m5)),
totpop_m6 = ifelse(totpop_m6 <= 0, NA, round(totpop_m6)))
```

6 Tables from the Replication of Lee et al. (2004)

Table A1: LMB's Balance Tests Using Extract Data Versus Our Harmonized Data

	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 et al. (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 percent mark. Column (6) includes a fourth order polynomial in Democratic vote share, which is interacted with the above-below 50 percent dummy. The unit of observation is the district-congress. Standard errors are clustered by district-decade. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: LMB's Balance Tests Using Extract Data Versus Our Harmonized Data (I)

	Difference in District Income Between Democrat and Republican Districts					
	(1)	(2)	(3)	(4)	(5)	(6)
Log income (M1)	-0.039***	-0.002	0.026*	0.033*	0.027	0.007
	(0.013)	(0.013)	(0.015)	(0.018)	(0.026)	(0.028)
Log income (M2)	-0.038***	-0.001	0.026*	0.033*	0.027	0.008
	(0.013)	(0.013)	(0.014)	(0.018)	(0.026)	(0.028)
Log income (M3)	-0.039***	-0.001	0.026*	0.033*	0.027	0.008
	(0.013)	(0.013)	(0.015)	(0.018)	(0.026)	(0.028)
Log income (M4)	-0.039***	-0.001	0.026*	0.033*	0.026	0.008
	(0.013)	(0.013)	(0.015)	(0.018)	(0.026)	(0.028)
Log income (M5)	-0.039***	-0.006	0.026*	0.037*	0.018	0.023
	(0.014)	(0.013)	(0.015)	(0.019)	(0.027)	(0.029)
Log income (M6)	-0.038***	-0.005	0.026*	0.038**	0.019	0.024
	(0.014)	(0.013)	(0.015)	(0.019)	(0.027)	(0.029)
Log income (LMB)	-0.087***	-0.037***	0.014	0.027	0.031	0.053*
	(0.013)	(0.013)	(0.014)	(0.018)	(0.026)	(0.029)
Bandwidth	All	$+/-$ 25	$+/-$ 10	$+/-$ 5	$+/-$ 2	Polynomial
Observations	13413	10229	4174	2072	810	13393

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

Table A3: LMB's Balance Tests Using Extract Data Versus Our Harmonized Data (II)

	Difference in District Urban Pop. Between Democrat and Republican Districts					
	(1)	(2)	(3)	(4)	(5)	(6)
% Urban (M1)	0.050^{***}	0.052^{***}	0.045^{***}	0.045^{***}	0.055^{***}	0.042^{*}
	(0.011)	(0.011)	(0.012)	(0.014)	(0.021)	(0.022)
% Urban (M2)	0.052^{***}	0.054^{***}	0.045^{***}	0.045^{***}	0.055^{***}	0.043^{*}
	(0.011)	(0.011)	(0.012)	(0.014)	(0.021)	(0.023)
% Urban (M3)	0.051^{***}	0.053^{***}	0.045^{***}	0.045^{***}	0.055^{***}	0.042^{*}
	(0.011)	(0.011)	(0.012)	(0.014)	(0.021)	(0.023)
% Urban (M4)	0.052^{***}	0.054^{***}	0.045^{***}	0.045^{***}	0.055^{***}	0.043^{*}
	(0.011)	(0.011)	(0.012)	(0.014)	(0.021)	(0.023)
% Urban (M5)	0.048^{***}	0.047^{***}	0.037^{***}	0.039^{***}	0.046^{**}	0.042^{*}
	(0.011)	(0.011)	(0.012)	(0.014)	(0.021)	(0.023)
% Urban (M6)	0.050^{***}	0.049^{***}	0.038^{***}	0.041^{***}	0.046^{**}	0.044^{*}
	(0.011)	(0.011)	(0.012)	(0.014)	(0.021)	(0.023)
% Urban (LMB)	0.070^{***}	0.066^{***}	0.054^{***}	0.054^{***}	0.056^{**}	0.053^{**}
	(0.011)	(0.011)	(0.013)	(0.015)	(0.023)	(0.025)
Bandwidth	All	+/− 25	+/− 10	+/− 5	+/− 2	Polynomial
Observations	13413	10229	4174	2072	810	13393

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

Table A4: LMB's Balance Tests Using Extract Data Versus Our Harmonized Data (III)

	Difference in District Blacks Between Democrat and Republican Districts					
	(1)	(2)	(3)	(4)	(5)	(6)
% Black (M1)	0.061***	0.037***	0.013***	0.007	-0.001	-0.006
	(0.005)	(0.004)	(0.005)	(0.006)	(0.009)	(0.010)
% Black (M2)	0.062***	0.037***	0.013***	0.007	-0.001	-0.005
	(0.005)	(0.004)	(0.005)	(0.006)	(0.009)	(0.010)
% Black (M3)	0.062***	0.037***	0.013***	0.007	-0.001	-0.006
	(0.005)	(0.004)	(0.005)	(0.006)	(0.009)	(0.010)
% Black (M4)	0.062***	0.038***	0.013***	0.007	-0.001	-0.006
	(0.005)	(0.004)	(0.005)	(0.006)	(0.009)	(0.010)
% Black (M5)	0.056***	0.033***	0.007	0.001	-0.006	-0.012
	(0.005)	(0.005)	(0.005)	(0.006)	(0.010)	(0.011)
% Black (M6)	0.057***	0.034***	0.008	0.001	-0.005	-0.012
	(0.005)	(0.004)	(0.005)	(0.006)	(0.010)	(0.011)
% Black (LMB)	0.083***	0.043***	0.014***	0.003	-0.003	-0.053***
	(0.006)	(0.005)	(0.005)	(0.006)	(0.009)	(0.012)
Bandwidth	All	$+/-$ 25	$+/-$ 10	$+/-$ 5	$+/-$ 2	Polynomial
Observations	13413	10229	4174	2072	810	13393

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

Table A5: LMB's Balance Tests Using Extract Data Versus Our Harmonized Data (IV)

	Difference in District Manufacturing Between Democrat and Republican Districts					
	(1)	(2)	(3)	(4)	(5)	(6)
% Manufacturing (M1)	-0.003	-0.000	0.004*	0.005*	0.004	0.002
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
% Manufacturing (M2)	-0.002	0.000	0.004*	0.005*	0.004	0.002
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
% Manufacturing (M3)	-0.002	0.000	0.004*	0.005*	0.004	0.002
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
% Manufacturing (M4)	-0.002	0.000	0.004*	0.005*	0.004	0.002
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
% Manufacturing (M5)	-0.002	-0.001	0.003	0.005	0.002	0.001
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
% Manufacturing (M6)	-0.002	-0.000	0.003	0.005*	0.002	0.001
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
% Manufacturing (LMB)	-0.002	0.000	0.004*	0.005*	0.004	0.003
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
Bandwidth	All	+/− 25	+/− 10	+/− 5	+/− 2	Polynomial
Observations	13413	10229	4174	2072	810	13393

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

Table A6: LMB's Balance Tests Using Extract Data Versus Our Harmonized Data (V)

	Difference in District Voters Between Democrat and Republican Districts					
	(1)	(2)	(3)	(4)	(5)	(6)
% Eligible to vote (M1)	0.004	0.008***	0.007**	0.006	-0.006	-0.013*
	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)	(0.007)
% Eligible to vote (M2)	0.004	0.009***	0.007**	0.006	-0.006	-0.013*
	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)	(0.007)
% Eligible to vote (M3)	0.004	0.009***	0.007**	0.006	-0.006	-0.012*
	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)	(0.007)
% Eligible to vote (M4)	0.004	0.009***	0.007**	0.006	-0.006	-0.012*
	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)	(0.007)
% Eligible to vote (M5)	0.005	0.008***	0.008**	0.006	-0.007	-0.011
	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)	(0.007)
% Eligible to vote (M6)	0.005*	0.009***	0.008**	0.006	-0.007	-0.011
	(0.003)	(0.003)	(0.004)	(0.004)	(0.007)	(0.007)
% Eligible to vote (LMB)	0.005*	0.011***	0.007**	0.007*	-0.003	-0.004
	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)	(0.006)
Bandwidth	All	+/− 25	+/− 10	+/− 5	+/− 2	Polynomial
Observations	13413	10229	4174	2072	810	13393

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

References

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse, "Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States," *Econometrica*, 2020, *88* (6), 2329–2368.

Eckert, Fabian, Andres Gvirtz, Jack Liang, and Michael Peters, "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790," *NBER Working Paper No. 26770*, 2020.

Fang, Yu and James W. Jawitz, "High-resolution reconstruction of the United States human population distribution, 1790 to 2010," *Scientific Data*, 2018, *5*, https://doi.org/10.1038/sdata.2018.67.

Haines, Michael, "Historical, Demographic, Economic, and Social Data: The United States, 1790-2002," *Inter-university Consortium for Political and Social Research [distributor], Ann Arbor, MI, 2010-05-21. https://doi.org/10.3886/ICPSR02896.v3*, 2010.

Hornbeck, Richard, "Barbed Wire: Property Rights and Agricultural Development," *Quarterly Journal of Economics*, 2010, *125* (2), 767–810.

Lee, David S., Enrico Moretti, and Matthew J. Butler, "Do Voters Affect or Elect Policies? Evidence from the U. S. House," *Quarterly Journal of Economics*, 2004, *119* (3), 807–859.

Lewis, Jeffrey B., Brandon DeVine, Lincoln Pritcher, and Kenneth C. Martis, *United States Congressional District Shapefiles*, 2021, *https://cdmaps.polisci.ucla.edu/ (Accessed on June 30, 2021).*

Leyk, Stefan, Johannes H. Uhl, Dylan S. Connor, Anna E. Braswell, Nathan Mietkiewicz, Jennifer K. Balch, and Myron Gutmann, "Two Centuries of Settlement and Urban Development in the United States," *Science Advances*, 2020, *6*, 1–12.

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles, "IPUMS National Historical Geographic Information System," *Version 15.0 [dataset]. Minneapolis, MN. DOI: http://doi.org/10.18128/D050.V15.0.*, 2020.

Uhl, Johannes H. and Stefan Leyk, "Historical built-up areas (BUA) - gridded surfaces for the U.S. from 1810 to 2015," *Harvard Dataverse*, 2020, *V1*, https://doi.org/10.7910/DVN/J6CYUJ.

_ and _ , "Historical built-up property records (BUPR) - gridded surfaces for the U.S. from 1810 to 2015," *Harvard Dataverse*, 2020, *V1*, https://doi.org/10.7910/DVN/YSWMDR.