C A G E

New area- and population-based geographic crosswalks for US counties and congressional districts, 1790-2020

CAGE working paper no. 588

October 2021

Andreas Ferrara
Patrick A. Testa
Liyang Zhou

WARWICK
THE UNIVERSITY OF WARWICK

UKRI Economic and Social Research Council

New Area- and Population-based Geographic Crosswalks for U.S. Counties and Congressional Districts, 1790-2020*

Andreas Ferrara† Patrick A. Testa‡ Liyang Zhou§

October 14, 2021

Abstract

A common problem in applied research involves harmonizing geographic units across time or different levels of aggregation. One approach is to use "crosswalks" that associate factors located within some "origin" unit to different "reference" units based on relative *areas*. We develop an alternative approach based on relative *population*, accounting for heterogeneities in urbanization within counties. We construct population-based crosswalks for 1790 through 2020, mapping county-level data across U.S. Censuses as well as from counties to congressional districts. Using official Census data for congressional districts, we show that population-based weights outperform area-based ones in terms of similarity to official data.

Keywords: boundary harmonization; geographic crosswalks; spatial population distribution.
JEL Codes: R12, C18, C59.

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

†University of Pittsburgh, Department of Economics. Email: a.ferrara@pitt.edu.
‡Tulane University, Department of Economics. Email: ptesta@tulane.edu.
§University of Pittsburgh, Department of Economics. Email: liz113@pitt.edu.

1 Introduction

It is common in economics and other social sciences to analyze data with a geospatial component.[1] Typically, data map onto geographic units that have changing boundaries over time, such as U.S. counties across Census years (Hornbeck, 2010). Meanwhile, data that are associated with different levels of spatial aggregation often need to be merged – for instance, when trying to combine county- and commuting zone-level information (e.g. Autor, Dorn, Hanson and Majlesi, 2020). If individual-level or other more local data are not available, researchers must rely on "crosswalks" to associate aggregate data across different units. Such crosswalks provide the researcher with means to disaggregate data associated with one spatial unit so that they may be re-aggregated within the boundaries of some different set of "reference" units.

A commonly-used approach to boundary "harmonization" is pioneered by Hornbeck (2010). First, the researcher chooses as the set of reference units that which is available at the highest level of aggregation. In the U.S. setting, for instance, newer counties are typically subsets of older, larger counties, allowing re-aggregation of counties backward in time. Then, to deal with cases in which unit boundaries do not neatly coincide, one can (i) intersect the boundaries of each set of "origin" units with those of the reference units, (ii) compute the area of overlap between the two relative to that of each given origin unit, and (iii) collapse relevant stock data associated with the origin unit within the reference boundary, using the share of overlap as weights. A key assumption underlying this process is that the factors measured by the aggregate data (e.g. population stocks) are *uniformly distributed* in space within the boundaries of the origin unit being harmonized. Recent work by Perlman (2021) and Eckert, Gvirtz, Liang and Peters (2020) has extended this approach to counties across all U.S. Census years and commuting zone delineations,[2] and a large number of papers in recent years have adopted this approach for the purposes of both intertemporal spatial analysis (Hornbeck and Naidu, 2014; Bazzi, Fiszbein and Gebresilasse, 2020; Calderon, Fouka and Tabellini, 2020; Ferrara and Testa, 2020) and spatial harmonization across different contemporaneous units (Eckert et al., 2020; Testa, 2021).

This paper makes two contributions to the method of spatial harmonization, with potential for broad application among urban economists, economic historians, political scientists, and other spatial researchers. First, we address concerns that the "uniformity assumption" underlying area-based weights may generate error in harmonized data when boundaries do not neatly coincide across origin and reference units (Hanlon and Heblich, 2020), such as when data are being merged across different levels of spatial aggregation. To do this, we introduce a procedure for generating *population-based* weights in the context of the conterminous U.S. between 1790 and 2020, based on new spatial models of population distribution by Fang and Jawitz (2018). These spatial models use population data from the U.S. Census alongside geographic

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

[2] For Perlman's crosswalks based on the uniformity assumption, see https://elisabethperlman.net/code.html.

and topographic features to approximate granular population distributions at the 1×1 kilometer grid cell level.[3] We use these maps to produce crosswalks that relax the uniformity assumption and identify where populations are more concentrated within counties. This is useful to the extent that harmonization first involves spatial disaggregation of county-level stock data. In particular, 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 important both for harmonizing population stock data as well as data for other economic factors, such as total income and number of college educated, that are likely to be correlated with urban density (i.e. due to agglomeration economies).[4] We use these new weights to extend previous county-to-country crosswalks across all Census years from 1790 (Hornbeck, 2010; Eckert et al., 2020). Our method is algorithmically similar to the procedure in Beddow and Pardey (2015), which uses information on the spatial distribution of production in U.S. as of 2000 to map historical county-level crop data to that year's boundaries.

Secondly, 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 (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 – and even 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 often associated with smaller CDs, an area-based weight is likely to underestimate the population of an urban CD and overestimate the population of a non-urban CD located within the same county. The more concentrated an urban agglomeration is relative to its county area (e.g. as in mountainous or marshland areas), the greater this bias will tend to be. Population-based weights help us overcome such bias.

In addition to producing these new weights and crosswalks, we also crosscheck their accuracy via an application. In particular, we replicate the CD-level data and the balance tests that underscore the regression discontinuity (RD) design used in Lee et al. (2004). These test for the exogeneity of the CD characteristics around the RD threshold, as the basis of their identification strategy. To measure CD characteristics, the authors importantly use official CD-level data from the "extract" versions of 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, reaffirming the identification strategy in Lee et al. (2004), data constructed using population-based weights consistently outperform

[3]The areal extents of urban areas are extrapolated backward in time, using scaling factors derived from urban population distributions in the year 2000.

[4]In other words, economic activity tends to locate where the people are located.

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 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. We focus on the construction of new county-to-congressional district (CD) crosswalks for the U.S. from 1790 to 2020, spanning the 1st through 116th U.S. Congresses, as harmonization across geospatial units 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.[5]

We use these methods to construct county-to-CD crosswalks 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.[6] Each crosswalk file includes four kinds of weights: (i) area-based (model 1, or M1); (ii) population-based (M2), with area divided into urban and rural areas; (iii) population-based (M3), with area divided into urban and rural areas after excluding non-inhabitable areas; and (iv) population-based (M4), with area divided into urban and rural areas after excluding non-inhabitable areas, with additional weighting for topographic suitability. M1 is equivalent in construction to existing area-based crosswalks (Eckert et al., 2020; Perlman, 2021), whereas M2-4 use models of population distribution from Fang and Jawitz (2018) at the 1×1 kilometer grid cell level.[7] We also construct county-to-county crosswalks for any two Censuses from 1790 to 2020, using both area-based weights and the three new population-based weights. Between the county-to-CD and county-to-county crosswalks, our crosswalks can be used to harmonize any county to any CD in U.S. history.

2.1 Constructing 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

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

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

[7] Two exceptions 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. For 2020, we use 2010 population distribution to construct population-based weights.

are then disaggregated into a set of sub-county units (henceforth "county-parts"), based on the CD in which they are located. Counties that lie wholly within a CD without intersecting its boundaries are their own and only county-part. Counties that are intersected by a single CD boundary are located partly in two CDs and thus have two county-parts. We then calculate the areas (in square meters) of all counties, all CDs, and all county-parts.[8] 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.

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. 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 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.

For other CDs, such as CD 8, this is only partly the case. That is, most CD-level data can be calculated by adding up the populations and sub-populations of whole counties, with one exception. For CD 8 in the state's northeast corner, which consists of 10 whole counties, this exception is Anoka County, of which a small portion – about 1/20th of the area of the whole county – is instead part of CD 5 alongside 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 would be estimated as having 10,950 residents.

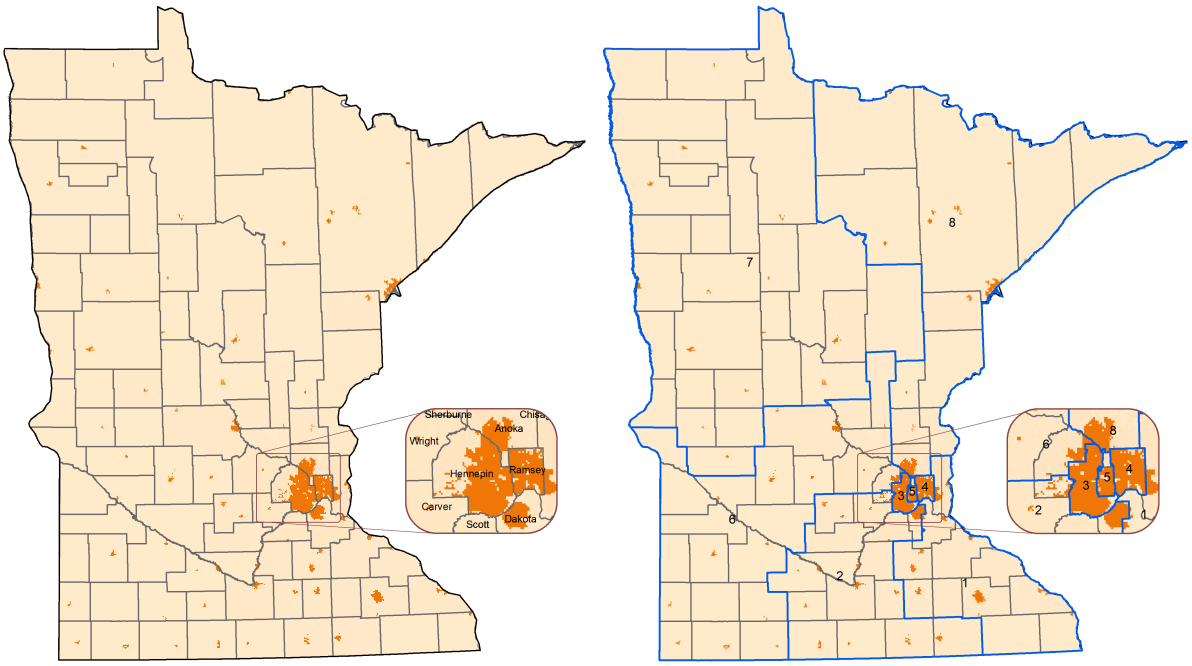
More complicated still is the process of harmonizing county-level data to CD 5's boundaries. Data values for CD 5 can be estimated by adding together a given area-weighted value from the remaining 1/20th of Anoka County with the area-weighted value of the part of Hennepin County that lies within CD 5. Note that Hennepin County is split between CDs 2, 3, 5, and 6. The part that lies in CD 5 is only about 1/10th of the county's total area. Hence,

[8] 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 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 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, Schroeder, Van Riper, Kugler and Ruggles (2020). CD shapefiles are from Lewis, DeVine, Pritcher and Martis (2021). Population distribution information for 1970 comes from M3 in Fang and Jawitz (2018).

under an area-based crosswalk, 1/10th of its population is allocated to CD 5. If every county in Minnesota had a population of 1000, CD 5 would have 150 residents: 50 from Anoka County and 100 from Hennepin County.

Note that there are potential drawbacks to using this area-based method when origin and reference unit boundaries do not neatly coincide, as is the case here. In particular, it only works under certain conditions on the distribution of population. To motivate this caveat, 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, while only about a tenth of the area of Hennepin County is 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 it – significantly underweighting this county-part, while overweighting all the others.

We will now discuss the theoretical underpinnings of the area-based weights used here and in prior area-based crosswalks, including conditions under which these weights are appropriate. We will then examine how to relax such conditions for settings in which they are not appropriate, using a set of novel *population-based weights*.

When is an area-based crosswalk appropriate?

Suppose a researcher is attempting to associate several county-level stock variables with congressional districts. For our area-based weights to be appropriate in settings where county and CD boundaries overlap, the following 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 standard stock variables at the county and sub-county (e.g. neighborhood) level, respectively, such as total population, total income, the total number of Spanish-speakers, etc. in their respective areas. Hence, when uniformity holds, a neighborhood's share of a given sub-population in a county is always equal to its share of the county's total area. Uniformity does not, however, mean that population is uniformly distributed *across* counties. Whenever counties neatly fall within CDs, our area-based crosswalks will capture sufficient population heterogeneity in space to accurately derive these stock variables at the CD level.

Now consider the following result:

Proposition. *Suppose all counties satisfy uniformity. Then our area-based crosswalk will accurately map county-level values to the congressional district level for all districts.*

Proof. See the Online Appendix. □

Given the ideal nature of this result, a researcher using area-based weights will want the uniformity assumption to either be as plausible as possible, or as irrelevant as possible. It might be plausible, for instance, in relatively low-density settings, such as farmland, with relatively homogeneous populations. And it will be less relevant a concern in settings in which harmonization is taking place from highly disaggregated data, or when the origin units lie neatly within the reference units, with little overlap in boundaries. For instance, 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 – for instance, due to urbanization or in the presence of agglomeration forces making the distribution of population uneven across space. To the extent that reference and origin unit boundaries overlap, area-based crosswalks will thus tend to generate error in the harmonization process whenever a county must be disaggregated, i.e. when a county lies in two or more CDs. This is because for each of a county's "county-parts" that is associated with a different CD, all stock data values as a share of the total county's are calculated as being equal to that county-part's share of the county's area under an area-based crosswalk. Yet when uniformity does not hold, a county-part's stock variables

may in reality be smaller or larger than their relative area, such as if it is more urban than the rest of the county. The more often county and CD boundaries do not coincide, the more such error is likely to occur and accrue. To address this, we also construct three population-based crosswalks in addition to the area-based crosswalk, which allow for heterogeneous population distribution within counties.

2.2 Constructing Population-based Crosswalks

We now relax the uniformity assumption, using information on historical within-county population distribution from Fang and Jawitz (2018). They provide population estimates at the 1×1 kilometer grid-cell level for three additional models: (i) area divided into urban and rural areas based on urban population being distributed around city centers according to a power law scaling relationship (model 2, or M2); (ii) a version of M2 that first excludes non-inhabitable areas, such as bodies of water (M3); and (iii) a version of M3 that applies additional weighting based on topographic suitability (M4).[9]

To estimate the spatial extent of urban areas for the conterminous United States over time, Fang and Jawitz (2018) use population distribution information for urban areas from the 2000 Census. They then extrapolate the size of the urban area to previous Census years, using the following power law scaling relationship,

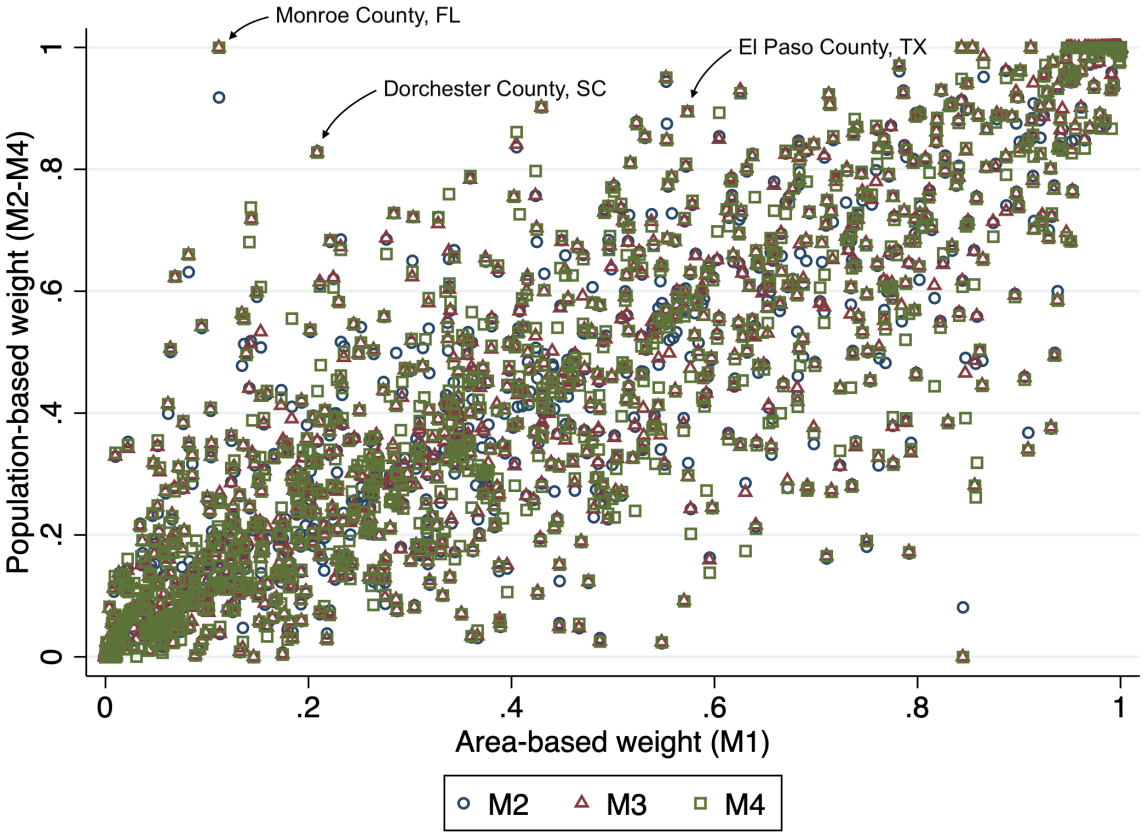
$$A_{U,\varphi} = \alpha_\delta P_{U,\varphi}^{\beta_\delta} \tag{1}$$

where $A_{U,\varphi}$ is the spatial extent of urban area and urban areas are indexed by φ in Census division δ, and where α_δ and β_δ are the coefficients of the power function, which are estimated based on the areas and populations of U.S. cities in 2000 and assumed to be constant over time. Using historical population data from the Census, Fang and Jawitz (2018) then estimate the historical areal extents of urban areas back to 1790. 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). For an in-depth description of these models and their underlying data, see the Online Appendix.

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 size*, using these population distribution maps from Fang and Jawitz (2018). These let us calculate for each Census year the total population of each county and of each county-part within that county that lies in a different CD (see Figure 1). As with our area-based crosswalk, the ratio of these populations provides a weight with which to multiply a county's stock data prior to

[9]Note that M2 still assumes some uniformity, within urban and rural areas; this is further relaxed by M3-4.

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) lack historical population distribution information.

its aggregation to the CD level.[10] Unlike our area-based crosswalk, however, relatively small county-parts in terms of area might receive a relatively large weight, for instance if they are associated with an urban area. Because CD boundaries are often associated with urban density, this occurs often across the set of counties.[11]

Such discrepancies between area- and population-based weights are shown in Figure 2, which relates weights from each of the three population-based based models to those from the area-based one for the 2010 Census and the 112th Congress. Although weights are highly correlated across models overall, some weights differ significantly. Take Dorchester County, SC, a suburban county that partially overlaps with the Charleston metropolitan area. As of 2011, nearly 80% of its area was in 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

[10]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.

[11]This may be particularly true in settings with greater partisan gerrymandering, in which voter characteristics are targeted to maximize party vote shares.

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 swamp and wetland areas such as Southern Louisiana and the Florida Peninsula.

2.3 Implementing the Crosswalks

Our crosswalks can be used to harmonize historical county boundaries to any other Census year, between 1790 and 2020. Our crosswalks can also be used to harmonize county boundaries to proximate CD boundaries. The latter crosswalks include 3 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 based on M1-4, except for Alaska and Hawaii, which only include M1. Note that between the county-to-CD and county-to-county crosswalks, our crosswalks can be used to harmonize any county to any CD in U.S. history.

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 data of interest is the percent of the population that was born in Mexico.

1. Get the county-level data for 1960 for two variables: (i) total population and (ii) total number of persons born in Mexico. It is critical to harmonize only county-level stock variables for weights to be appropriate. If source data are shares or average outcomes, one should transform the variable first, e.g. by multiplying by total population.

2. Given some set of county identifiers (e.g. FIPS or NHGIS codes), merge the 1960 county file with the 1960 to 88th Congress crosswalk file. This expands the set of counties into the full set of county-parts, based on the CDs they are associated with.

3. Take stock 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. This will transform the stock variables into measures proportional to those weights. Weights may differ across the four 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 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 design used in Lee et al. (2004). These test for the exogeneity of the CD characteristics around the tied-election threshold as the basis of their identification strategy. To measure CD characteristics, the authors importantly use official CD-level data from the "extract" versions of 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 these balance tests, are replicated when CD characteristic data are harmonized from county-level data, as well as whether this differs across our four crosswalk models.

We begin by using our crosswalks to construct CD-level stock data from the county-level Census data, with which to compare to the official extract 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 population, (iv) Black population, (v) number of manufacturing workers, and (vi) number of eligible voters.[12] These reconstructions compare favorably across all four of our crosswalk models to the extract data, as gauged by their correlations with the latter, as shown in Figure 3. In general, however, M1 always performs worse than our population-based crosswalks. On average, the correlation between M2-4 and the official data is 0.9 whereas it is 0.76 for the M1 model. This means that that population-based data improve the correlation with the official data by almost 20% relative to the area-based data. Meanwhile, among the three population-based models, none clearly or consistently outperform the others.

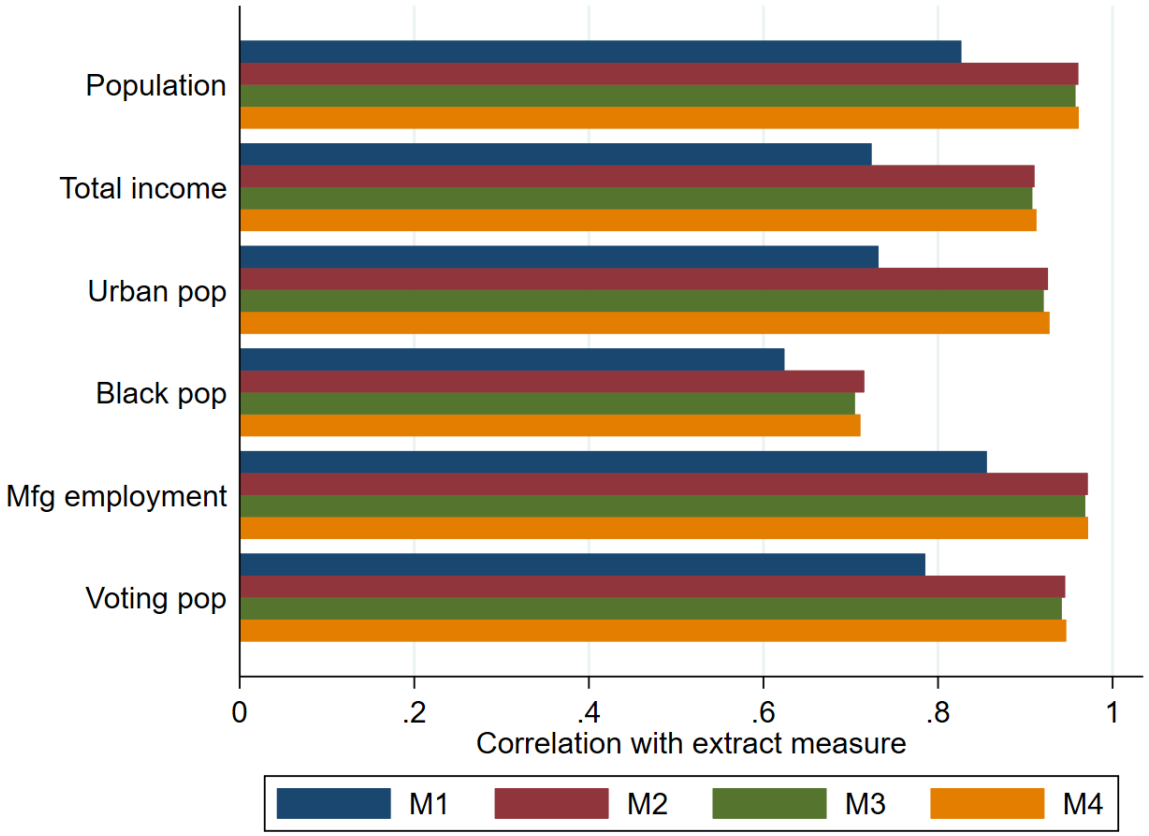
Of the six variables we reconstruct, the total population and manufacturing population data are closest to the official extract data, while the number of Blacks is the most different. This makes sense if you consider where Blacks tend to live in the U.S. In regions like the Midwest and Northeast, such as in states like Illinois, Michigan, and Maryland, Blacks tend to live disproportionately in 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 data whether a particular variable is appropriate, given its spatial distribution relative to a county's area or overall population.[13]

On the other hand, there are clear upsides to our approach, i.e. to using county-level data harmonized to the CD level. Extract data such as that 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. As a result, such datasets often associate CD socioeconomic characteristics

[12] 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.

[13] For instance, 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).

Figure 3: Comparison of Harmonized Data Versus Official CD-Level Extract Data



Note: Figure compares harmonized CD-level data generated by our four crosswalks to official CD extract data, as featured in Lee et al. (2004), from the U.S. Censuses 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 extract data are 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.

with electoral outcomes that are based on different CD boundaries. They are also limited to a relatively small set of Census variables, whereas spatial researchers often deal with novel county-level data constructed from historical data not found in the Census. 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 replicate the balance tests using their official data and code. These are done without issue; as in their paper, % urban and % Black show slight but statistically significant discontinuities across multiple specifications, but most observable characteristics suggest few differences between Democratic and Republican CDs around the 50% threshold. We then do the same using our four models. Our balance tests, shown in Tables A1-5 in the Online Appendix, reaffirm the identification strategy in Lee et al. (2004). If anything, we find fewer discontinuities for narrow-bandwidth balance tests, while reaffirming the slight discontinuity for % urban. As an example, the balancing test for their population data is shown in Table 1, alongside our four models.

Table 1: 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)	-92262.930***	-72968.323***	-23473.905**	-24417.836*	-34212.160	-12495.686
	(12117.088)	(12874.551)	(11741.629)	(12977.439)	(22510.551)	(21968.564)
Total pop (M2)	-36638.876***	-18049.286***	-3286.556	-3727.587	-1255.204	-60.112
	(5984.388)	(5947.650)	(6254.601)	(7972.550)	(12973.779)	(13870.008)
Total pop (M3)	-37867.347***	-18590.271***	-3698.921	-4059.751	-1240.913	238.748
	(6215.897)	(6157.839)	(6345.093)	(8065.705)	(13139.512)	(14192.789)
Total pop (M4)	-31610.827***	-14373.230**	-2179.311	-3198.215	1262.041	3522.363
	(5944.258)	(5908.087)	(5955.189)	(7710.239)	(12849.999)	(13725.372)
Total pop (LMB)	-1817.582	3019.938	4961.497	3211.090	8640.547	2007.957
	(3517.336)	(3723.368)	(4562.725)	(5524.225)	(8427.041)	(9258.118)
Bandwidth	All	+/− 25	+/− 10	+/− 5	+/− 2	Polynomial
Observations	13231	10065	4086	2030	794	13211

Note: Each row features estimates from a different harmonization model, except for row (5), which uses data and code from Lee et al. (2004). 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 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4 Conclusions

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 different contemporaneous units. Existing approaches have often used the relative area of overall between different units to generate and apply weights to stock data for origin units, for the purposes of deaggregating and re-aggregating them to some reference unit. These approaches assume a uniform distribution of factors within origin units. In this paper, we develop an alternative approach based on historical models of U.S. population distribution by Fang and Jawitz (2018), in which weights are instead based on relative *population*. This mitigates issues present when economic factors are unevenly distributed within counties.

We use these methods to produce a set of novel crosswalks, which relax the uniformity assumption and apply 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 often 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 crosswalks reaffirm their identification strategy, data constructed using population-based weights consistently outperform area-based ones in terms of similarity to official data. We hope these methods and the crosswalks produced with them will be of value to spatial

researchers across the social sciences, for whom novel historical data are often 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, 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.

Calderon, Alvaro, Vasiliki Fouka, and Marco Tabellini, "Racial Diversity, Electoral Preferences, and the Supply of Policy: The Great Migration and Civil Rights," *Harvard Business School BGIE Unit Working Paper No. 20-017*, 2020.

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, "Resource Blessing? Oil, Risk, and Religious Communities as Social Insurance in the U.S. South," *CAGE Working Paper No. 513*, 2020.

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.

Hanlon, Walker W. and Stephan Heblich, "History and Urban Economics," *NBER Working Paper No. 27850*, 2020, *125.*

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.

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).*

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.

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

Perlman, Elisabeth, "Tools for Harmonizing County Boundaries," *http://elisabethperlman.net/code.*

html. Accessed on May 18, 2021, 2021.

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§

October 14, 2021

Table of contents

*The crosswalks, teaching material, and code and data for the replication exercise can be downloaded at https://doi.org/10.3886/E150101

†University of Pittsburgh, Department of Economics. Email: a.ferrara@pitt.edu.
‡Tulane University, Department of Economics. Email: ptesta@tulane.edu.
§University of Pittsburgh, Department of Economics. Email: liz113@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.

Population distribution data come from Fang and Jawitz (2018). These are described in greater detail in the next section. We use Census 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.

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 Census county-level 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 Census, which is separately sourced from NHGIS.

2 Overview of population distribution 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),

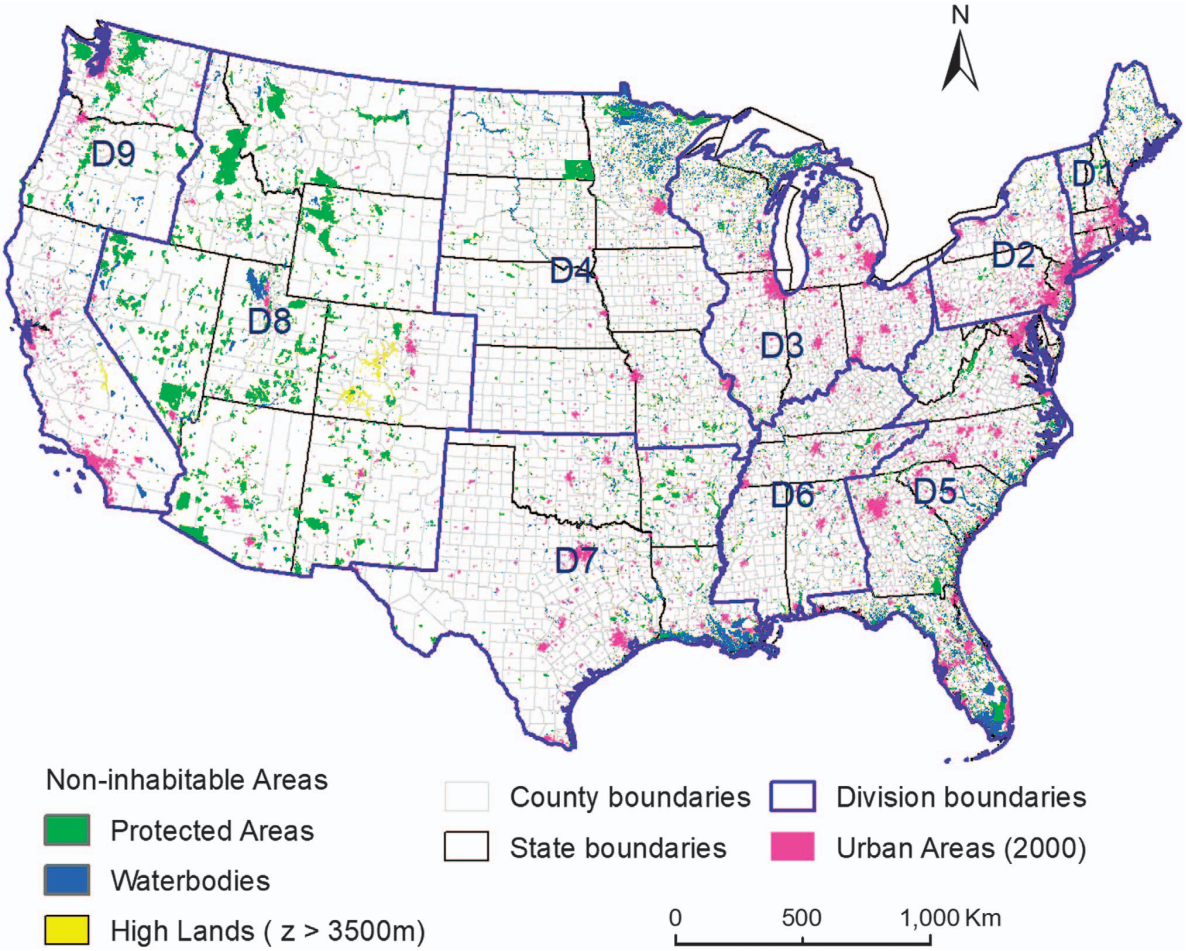
2. Data on water bodies from the National Hydrography Dataset (http://nhd.usgs.gov/) including lakes, ponds, marsh and swamp land, and other minor water bodies as of 2001,

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 Census from 1900 onward. An illustration of the spatial variation of their data for the year 2000 is shown in Figure A1:

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 Census divisions. Source: Figure 1 in (Fang and Jawitz, 2018, p. 3).

Defining the spatial extent of urban areas

Fang and Jawitz (2018) measure the areal extent of 3,610 urban areas using the 2000 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 Census 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 (M1) corresponds to the same underlying data used in Hornbeck (2010) and Perlman (2021). This approach assumes a 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 (M2) 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 (M3) also does this but first excludes non-inhabitable areas, as described above. The fourth model (M4) extends M3 by also multiplying the population raster by topographic suitability weights. We do not discuss or utilize the fifth model (M5), in which the authors also incorporate 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] Figure A2 compares these models for the year 2000.

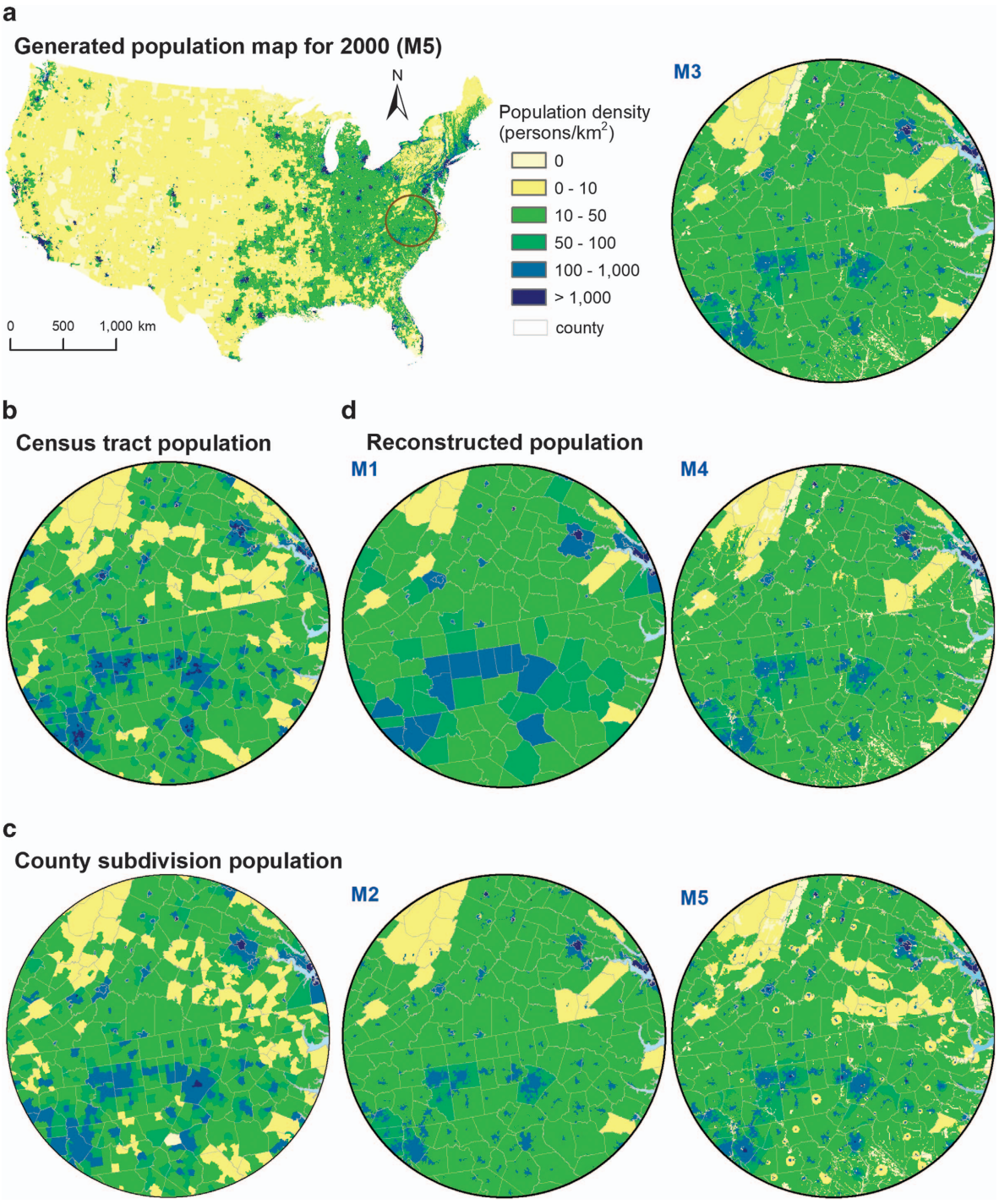
Alternative models for 1960

Because digital urban population data are missing for the 1960 Census, Fang and Jawitz (2018) do not develop models for that year. In lieu of this, and in the interest of completeness, we still

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

[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.

Figure A2: Comparison of estimated population distributions in 2000



Note: Map showing the generated population distribution based on model M5 in Fang and Jawitz (2018) (panel a), the true Census tract and county subdivision populations (panel b and c, respectively) for a chosen area in the South Atlantic region. Panel d shows the reconstructed population allocation based on models M1 to M5. Source: Figure 3 in (Fang and Jawitz, 2018, p. 9).

construct crosswalks based on 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 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.

Modeling population distribution for 2020

Given that Fang and Jawitz (2018) predates the 2020 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 Construction of our geographic crosswalks

This paper makes two 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. We begin by extending existing 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 historical population distribution models (M2-M4) from Fang and Jawitz (2018). These capture heterogeneities within the population distribution of each county being harmonized, in which more or less of its population may be located in some "parts" relative to others. Given that harmonization frequently involves spatial disaggregation of counties before re-aggregation to the boundaries of some "reference year," this better ensures that initial county stock variables will be properly weighted in the disaggregation process prior to being re-aggregated to different reference year 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. When used in conjunction with our county-to-county crosswalks, any county can be matched to any CD at any point in time. 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 again renders the historical population distribution models (M2-M4) from Fang and Jawitz (2018) quite important. We will now describe this process of disaggregation and re-aggregation, used to generate our county-to-county and county-to-CD crosswalks.

3.1 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.

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:

Proof of Proposition

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. □

3.2 Population-based crosswalks

We also construct population-based crosswalks, which rely on a relaxation of the uniformity assumption. To do this, we use information on historical within-county population distribution from Fang and Jawitz (2018). They provide population estimates at the 1×1 kilometer grid-cell level for three additional models: (i) area divided into urban and rural areas based on urban population being distributed around city centers according to a scaling function (model 2, or M2); (ii) a version of M2 that first excludes non-inhabitable areas, such as bodies of water (M3); and (iii) a version of M3 that applies additional weighting based on topographic suitability (M4). 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. Fang and Jawitz's (2018) population distribution raster maps let us calculate the total population of each county polygon for each Census year, as well as approximate population counts for each county-part polygon within a county that lies in a different CD. As with our area-based crosswalk, the population ratio of county-part to county provides a weight with which to multiply a county's stock data prior to its aggregation to the CD level.

3.3 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. After providing these examples, we will discuss step-by-step how to apply these weights and harmonize county boundaries to those of a contemporaneous CD, including relevant samples of Stata code.

3.3.1 Guide on 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 within each county-part, based on 1×1 raster cells from a given model from Fang and Jawitz (2018) for the county census year. These produce three separate tables, one each for M2, M3, and M4, with the same county-part identifiers used in the intersection output table. We then export all four 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 and intersection output tables. Total populations for each county for M2, M3, and M4, based on Fang and Jawitz (2018), are calculated by summing the populations of county-parts within each given county.

To generate weights for area-based crosswalks, we calculate the fraction of county-part to county area for each county-part. To generate weights for each population-based crosswalks, estimated county-part populations are divided by the total county population for each model. Miscellaneous 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.

4 Step-by-step guide on applying the crosswalks in Stata

Below we provide a simple example using Stata 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                      Number 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. Here we do this using the weights from all four models M1-4.

```
. * 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
. }
(0 real changes made)
(1 missing value generated)
(35 missing values generated)
(35 missing values generated)
(35 missing values generated)
(0 real changes made)
(1 missing value generated)
(35 missing values generated)
(35 missing values generated)
(35 missing values generated)
```

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)´ {
  1.      replace `v´ = . if(`v´<=0)
  2.      replace `v´ = round(`v´)
. }
(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)
(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)
```

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

5 Tables from our replication of Lee et al. (2004)

Table A1: 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)	(0.014)	(0.018)	(0.026)	(0.028)
Log income (M4)	-0.038***	-0.001	0.026*	0.033*	0.026	0.008
	(0.013)	(0.013)	(0.015)	(0.018)	(0.026)	(0.028)
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 (5), which uses data and code from Lee et al. (2004). 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 A2: 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)	0.052***	0.054***	0.045***	0.045***	0.055***	0.043*
	(0.011)	(0.011)	(0.012)	(0.014)	(0.021)	(0.023)
% Urban (M4)	0.053***	0.054***	0.045***	0.045***	0.055***	0.043*
	(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 (5), which uses data and code from Lee et al. (2004). 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 (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***	0.038***	0.013***	0.007	-0.001	-0.005
	(0.005)	(0.004)	(0.005)	(0.006)	(0.009)	(0.010)
% Black (M3)	0.062***	0.038***	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	-0.005
	(0.005)	(0.004)	(0.005)	(0.006)	(0.009)	(0.010)
% 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 (5), which uses data and code from Lee et al. (2004). 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 (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 (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 (5), which uses data and code from Lee et al. (2004). 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 (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	-0.013*
	(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	-0.013*
	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)	(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 (5), which uses data and code from Lee et al. (2004). 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

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).*

Perlman, Elisabeth, "Tools for Harmonizing County Boundaries," *http://elisabethperlman.net/code.html. Accessed on May 18, 2021*, 2021.