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

Andreas Ferrara† Patrick A. Testa‡ Liyang Zhou§

October 24, 2023

Abstract

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

Keywords: historical research; boundary harmonization; geographic crosswalks; spatial distribution of economic activity.
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 for social scientists to analyze data with a geospatial component.[1] When doing so, multiple 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 provide the researcher with means to disaggregate data associated with some *origin* spatial unit so that they may be re-aggregated within the boundaries of the *reference* unit on which the analysis focuses.

A typical approach to this boundary harmonization process was pioneered by 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 shares of C_2^{80}'s factors that lie in C^{70}, 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 in recent years 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, 2022; Ferrara and Testa, Forthcoming) and spatial harmonization across different contemporaneous units (Eckert, Gvirtz, Liang and Peters, 2020; Testa, 2021; Bazzi, Ferrara, Fiszbein, Pearson and Testa, 2022).

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 concerns that the uniformity assumption underlying area-based weights may generate error in harmonized data, to the extent that boundaries do not neatly coincide across origin and reference units (Hanlon and Heblich, 2020). To do this, we develop a procedure for generating a set of *population-based* weights in the context of the conterminous U.S. between 1790 and 2020, based on new spatial models of historical sub-county population distribution by Fang and Jawitz (2018) and Leyk, Uhl, Connor, Braswell, Mietkiewicz, Balch and Gutmann (2020). We use these 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 urban 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-country crosswalks across all U.S. census years (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 the U.S. as of 2000 to map historical county-level crop data to that year's boundaries. Our resource is complementary to the work of Berkes, Karger and Nencka (2021)—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, 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 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 using data that was collected at the CD level and compare it with data generated from crosswalked county-level information. For this purpose, we replicate the CD-level data and the balance tests that motivate the regression discontinuity (RD) design used 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, reaffirming the identification strategy 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. 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. 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]

We construct county-to-CD crosswalks based on: (i) the nearest census year, relative to the starting

[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.

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.
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 et al. (2020).

We also construct county-to-county crosswalks for any two 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 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 entail a process of spatial disaggregation and re-aggregation. For our county-to-CD crosswalks, this 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. 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.[5] 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 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

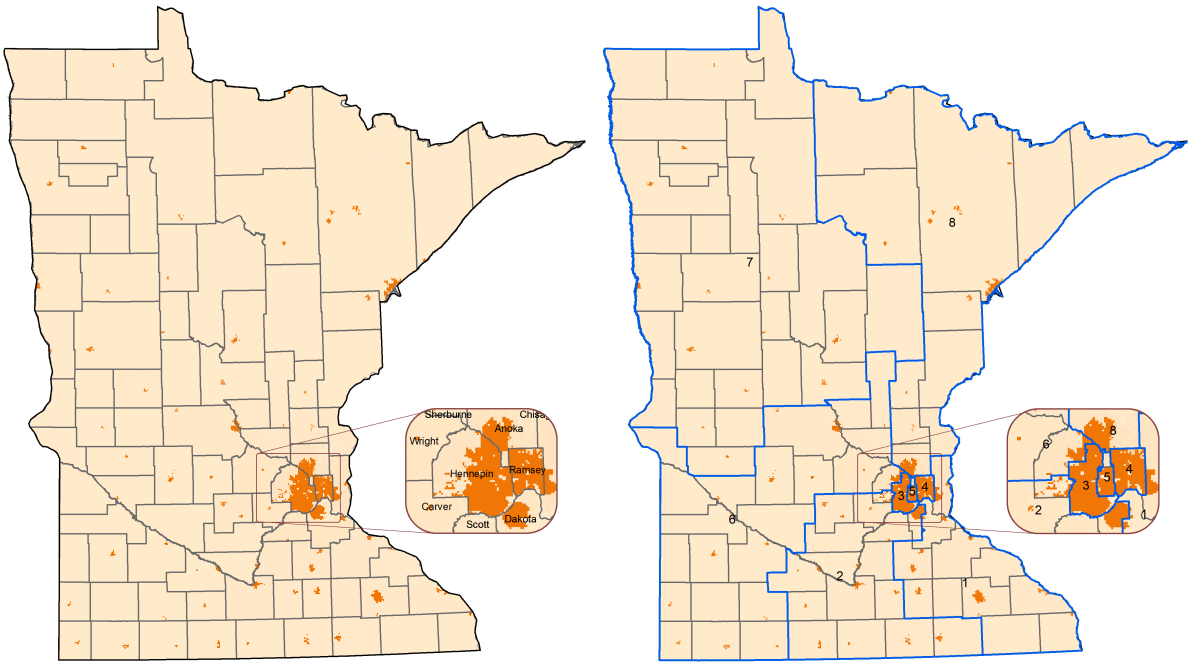
[4] For example, under the first approach, counties from the 1800 U.S. Census are harmonized to CDs for the 4th through 8th Congresses, spanning 1795 through 1804; under the second approach, counties from the 1800 U.S. Census are harmonized to CDs for the 7th through 11th Congresses, spanning 1801 through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

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, 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 the 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 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. Uniformity does not, however, mean that population is uniformly distributed *across* counties. Whenever counties neatly fall within CDs, 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 the 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 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 relative to ground-truth data 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 its 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 a set of 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 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 et al. (2020) (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 a power law scaling relationship;[6] (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. LU, in contrast, 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 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.[7] 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). To estimate the spatial extent of urban areas for the conterminous United States over time, FJ use population distribution information for urban areas from the 2000 U.S. Census. To the extent that historical spatial data do not exist, FJ then extrapolate the size of the urban area to previous census years, using the following power law scaling relationship,

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

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

[7] 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. 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.

where $A_{U,\varphi}$ is the spatial extent of an urban area, where urban areas are indexed by φ in U.S. Census Bureau 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, FJ 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).

While all of FJ's models of sub-county population rely on this assumption, a subset also make adjustments for the presence of non-inhabitable areas and topographic suitabilities—the basis for our weights M3 and M4, respectively. For an in-depth description of these models and the data used to construct them, see the Online Appendix. For more discussion of FJ's assumptions and potential drawbacks, see Section 4.

Describing the Spatial Models in Leyk et al. (2020). 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. On that basis, property records 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 a more in-depth description of the LU models and their underlying data, see the Online Appendix. For more discussion and model comparisons, see Section 4.

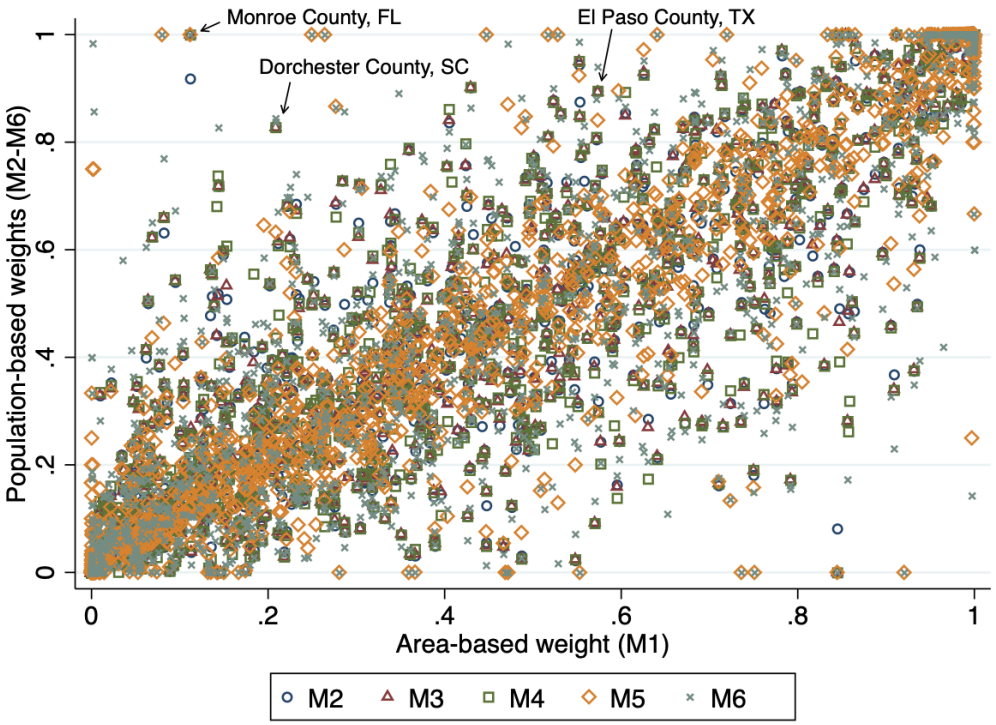
Constructing the Crosswalks. In order to relax the uniformity assumption, our population-based crosswalks no longer base the disaggregation of county-level data on relative area but rather on relative *population*, using these models of historical sub-county population distribution from FJ and LU. The resultant maps allow us to calculate for each census year the total population (or property-based proxy) count within each county, as well as for each county-part within that county that lies in a different CD (see Figure 1). We do this by summing the grid cell values within those respective polygons, using GIS software. As with our area-based crosswalk, the *ratio* of the latter count over the former provides a weight with which to multiply a county's relevant stock data prior to its aggregation to the CD level.[8] Unlike our 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. Because CD boundaries are often associated with urban density, this occurs often across the set of counties.[9]

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

[8] 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.

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

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 et al. (2020) lack historical population distribution information.

U.S. 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 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. Whether using area- or population-based crosswalks, one should choose as the set of reference counties that which is available at the highest level of aggregation.[10] Our crosswalks can also be used to harmonize county boundaries to 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)

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

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

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

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

3. Take 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 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.

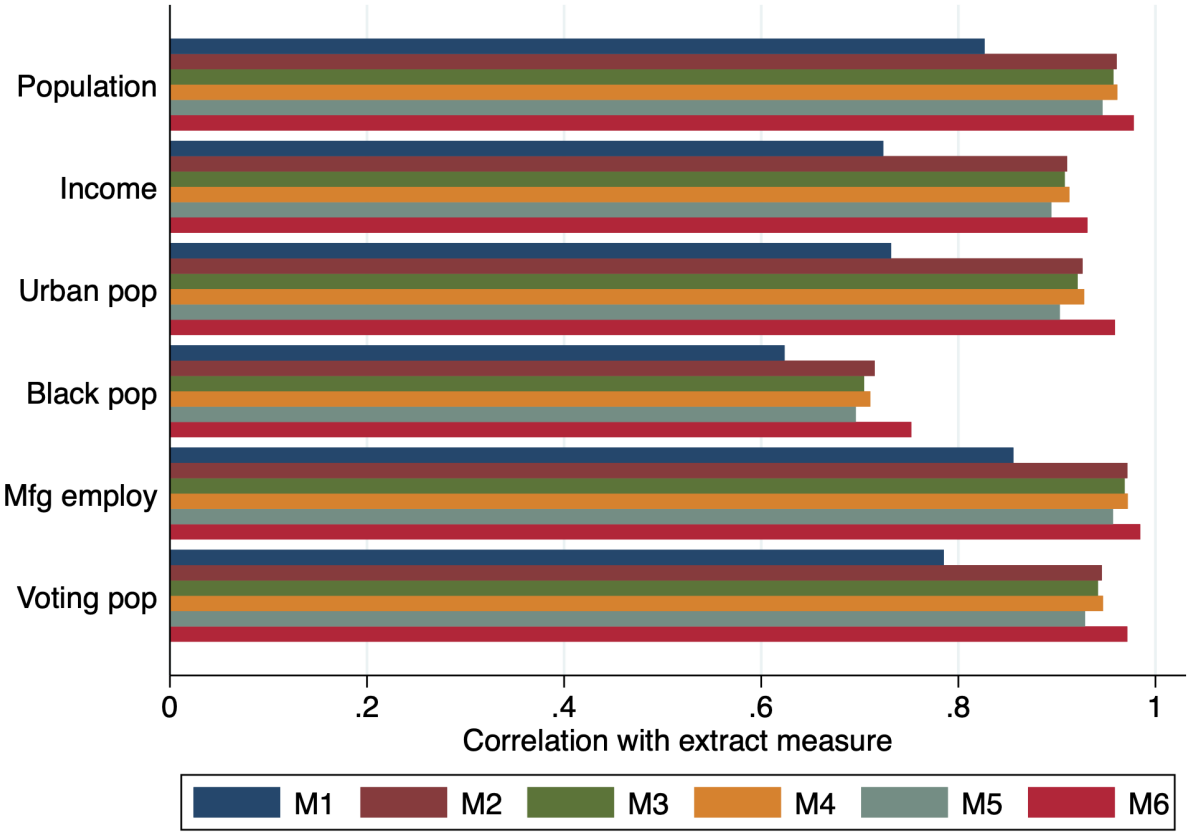
See the Online Appendix for sample Stata code demonstrating this process.

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 six crosswalk weighting 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

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



Note: Figure compares harmonized CD-level data generated by our six crosswalk weights to official CD extract 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 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.

eligible voters.[11] These reconstructions compare favorably across all six of our weighting 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 correlations between the data values generated using M2–M6 and the official data are around 0.9, whereas the correlation is about 0.76 for M1. 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 five population-based models, none clearly or consistently outperform the others, with the possible exception of M6—though we will discussion the limitations and inherent tradeoffs of using the various models more in Section 4.

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

[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.

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)	-37081.282***	-18546.004***	-3286.556	-3727.587	-1255.204	-336.524
	(5975.745)	(5935.772)	(6254.601)	(7972.550)	(12973.779)	(13872.485)
Total pop (M3)	-38286.800***	-19051.516***	-3706.085	-4059.751	-1240.913	13.916
	(6224.479)	(6156.620)	(6348.149)	(8065.705)	(13139.512)	(14200.134)
Total pop (M4)	-32030.089***	-14838.958**	-2192.660	-3198.215	1262.041	3319.982
	(5950.766)	(5905.531)	(5958.140)	(7710.239)	(12849.999)	(13732.539)
Total pop (M5)	-64661.918***	-47055.338***	-16416.888**	-11627.902	-17519.314	-2367.403
	(7097.914)	(7346.667)	(7778.517)	(9311.340)	(15465.836)	(15463.106)
Total pop (M6)	-20484.273***	-7429.907*	-1483.718	-729.554	2145.127	9868.712
	(4094.468)	(4192.568)	(5393.877)	(7358.480)	(13164.526)	(11854.139)
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 (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 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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, 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. 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 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 six weighting models. Our balance tests, shown in Tables A1–A5 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 six models. Among the six weighting models, estimates using M4 and M6 are closest to the ground-truth ones in the final column, while M1 and M5 are the furthest, mirroring Figure 3.

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 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 relative areas. To the extent that the spatial distribution of origin data often varies with urban density, we argue that our population-based crosswalks constitute an important improvement over these existing practices. Second, we are by no means claiming that our population-based crosswalks are *always* preferred over other approaches. We now address some limitations of our population-based crosswalks.

4.1 Limitations of Crosswalks

Error in harmonized data stems 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 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 would 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).

When Should You Use FJ- Versus LU-Based Weights? 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. For instance, 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. (2021) places 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. For instance, 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.

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. For instance, the ZTRAX property database from which the LU maps are constructed are increasingly unlikely to have property records in 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.[12] 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.[13]

Despite these caveats, it is reassuring that all of these population-based approaches outperform the area-based approach above in Figure 3, while being quite similar to each other in terms of accuracy. Yet because of these limitations, to the extent that population-based crosswalks are both needed and appropriate given the factors of study, we recommend using M2–M4 for 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, to

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

[13] For example, 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.

the extent that 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 use the relative area of overlap between different units to generate and apply weights to stock data for origin units, for the purposes of disaggregating and re-aggregating them to some reference unit. These approaches 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 et al. (2020), 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 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 crosswalks 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, Andreas Ferrara, Martin Fiszbein, Thomas Pearson, and Patrick A. Testa, "The Other Great Migration: Southern Whites and the New Right," *NBER Working Paper No. 29506*, 2022.

_ , 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," *Working Paper*, 2021.

Calderon, Alvaro, Vasiliki Fouka, and Marco Tabellini, "Racial Diversity, Electoral Preferences, and the Supply of Policy: The Great Migration and Civil Rights," *Review of Economic Studies*, 2022.

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*, Forthcoming.

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.

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, 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.

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

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