A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790[*]

Fabian Eckert[†]   Andrés Gvirtz[‡]   Jack Liang[§]   Michael Peters[¶]

February 2020

Abstract

Empirical researchers often have to map data provided for a "reporting" spatial unit, say counties in 1900, to a "reference" one, say, counties in 2010. We discuss a general method to create such *crosswalks*: computing the share of the area of each reporting unit nested in a given reference unit. Using these shares, data can be re-aggregated from the reporting to the reference units. We apply the method to construct a crosswalk for mapping US county-level data from 1790 to 2010 to present-day county or commuting zone delineations. We also provide the code to generate other crosswalks given maps of reporting and reference units.

Keywords: Regional Analysis, Regional Economic Activity, Regional and Urban History
JEL Codes: A1, N11

[†]Princeton University; fabian.eckert@princeton.edu
[‡]University of Cambridge; ag2038@cam.ac.uk
[§]Yale University; jack.liang@yale.edu
[¶]Yale University & NBER; m.peters@yale.edu

1. Introduction

Many social scientists work with data on regional outcomes, e.g., wages at the county level. However, data sources often report information on different levels of geographical aggregation, making the construction of unified data sets difficult. To map data between different geographies, *crosswalks* are needed; sometimes within and sometimes across years. For example, within a year, one data set may report school funding at the school district level, the other education outcomes for US Census tracts. Across years, 1940 population data may only be available for historical county delineations, while 2010 population numbers are available only for present-day delineations. In this paper, we describe a general method to create geographical crosswalks for both scenarios, and implement it in a simple Python routine.

To construct a geographical crosswalk using our method, two maps in the "shapefile" format are required: one for the spatial units for which data is reported (the "reporting" units or map) and one for the spatial units to which the data is to be re-aggregated (the "reference" units or map).[1] Given these two maps, our code calculates a set of "weights" that constitute the crosswalk. Users can then combine these weights with data available for the reporting units, and re-aggregate them to the reference units.

We apply our method to create crosswalks across time for two geographies frequently used in economic research: counties and commuting zones (as delineated by Tolbert and Sizer (1996)).[2] Our crosswalks can be used to map US county-level data from 1790-2010 to one of three alternative reference units: 1990 county boundaries, 2010 county boundaries, or 1990 commuting zones. We choose counties to showcase our method since they are the most detailed geographical unit reported in every Census since 1790. The crosswalk allows researchers to construct a county-level data set with consistent geographical units for the entire period of existence of the United States. At https://www.fpeckert.me/eglp, we provide code for the general method, replication code for the construction of our county and commuting zones crosswalks, and the crosswalks themselves.

To understand the idea behind our crosswalk, suppose one wants to display data available for a reporting map on the geographical delineations of a reference map. The two maps cover the same territory (e.g., the continental United States), but contain different partitions of it (e.g., counties from different time periods or Census tracts and states). Our method overlays the two maps using "Geographical Information Software" (GIS). The intersection of the two maps generates a new map consisting of smaller mutually exclusive spatial "subunits." These subunits can be combined to form the original partitions in *either* the reporting or the reference map.

The areas of these subunits can be used to construct weights to re-aggregate data provided for the

---

[1]Shapefiles are a special data format for maps. Shapefiles contain a collection of polygons oriented in space. Shapefile editors, e.g., ArcGIS or QGIS, can open shapefiles and display the polygons contained in them on the map of the world.

[2]For examples of recent work at this geographical resolution, see Autor and Dorn (2013), Autor et al. (2013), or Acemoglu and Restrepo (2017).

spatial units of the reporting map to the spatial units of the reference map. The collection of these weights constitutes our crosswalk. Weights are always associated with spatial units in both the reporting and reference map. To map stock data (e.g., population or payroll) the appropriate weights are simply the shares of the subunits in the total land area of their spatial unit in the *reporting* map. The underlying assumption is that the variable of interest is distributed proportionally to area size across the subunits contained within each spatial unit of the *reporting* map. To use the crosswalk, one assigns fractions of the variable of interest to the different subunits according to their weights and aggregates these subunits to the spatial units in the reference map.[3]

We apply our method to US county delineations between 1790 and 2010.[4] A key source for historical information on the US economy is the National Historical Geographic Information System (NHGIS) website (see Manson et al. (2017)). It provides a host of information (e.g., data on population size, the housing stock, agricultural land, and a variety of other economic variables) for various geographical units from 1790 to the present. However, NHGIS and similar sources report data with geographical identifiers current at the time of data collection. In the case of counties, enormous changes in their number and geographical delineations occurred between 1790 and today. Even counties with the same names may have very different shapes and locations in 1790 and 2010. Geographical changes are mostly due to the territorial expansion of the United States and population growth that led existing counties to split.[5] As a result, data for different decades cannot simply be combined into a decadal panel. For a longitudinal study, researchers require a crosswalk that maps all data to a common set of geographical units. We discuss intricacies specific to the creation of a county crosswalk for the United States and use our method to construct such a crosswalk that maps all counties since 1790 to present-day counties or commuting zones.

The online data repository associated with this project (https://www.fpeckert.me/eglp) contains Python code to intersect two different maps of the same geography (e.g., a county and a Census tract map of the United States) and calculate the collection of weights that constitute a crosswalk between them. Hence, the code is not specific to our application, but can easily be amended to construct crosswalks for arbitrary geographies. In Appendix B, we provide step-by-step instructions for researchers to use our code. The repository also offers our crosswalk for US counties since 1790

---

[3]While we intend our crosswalk to be primarily used with data on stocks, weights to map averages (e.g., average wages) can also be constructed. The correct weights, in this case, are found by computing for each spatial unit in the *reference* map, the share of each of its subunits in its total land area. In the crosswalk files provided on our website, we include the weights for stock data. We also provide the areas of all subunits which can be used to construct weights for data on averages.

[4]The crosswalk described in this paper was constructed as part of the empirical analysis in Eckert and Peters (2018). Eckert and Peters (2018) aggregate historical county-level data to the level of 1990 commuting zones to study the joint movement of US workers across sectors and labor markets between since 1880.

[5]Territorial expansion led to the creation of more and more administrative spatial units. Population growth caused existing administrative spatial units to split to maintain a similar population distribution across administrative entities. A host of other factors drove changes in county boundaries over the decades. To borrow an example from Horan and Hargis (1995), William County, North Dakota, was annexed to Mercer and Stark Counties in 1899 and William County ceased to exist. Within the year, but in a different section of the state, a "new" William County was established. As a result, the same county name appears in two successive decennial Censuses, but refers to entirely different geographical areas.

and replication files that document its construction. Lastly, it includes example code that applies our crosswalk to map 1790 county-level population data to 2010 county delineations.

We are not the first to construct a crosswalk from historical to present-day counties. Horan and Hargis (1995) produced a "County Longitudinal Template" that allows for an inter-temporal comparison of fixed county groups between 1840 and 1990. Rather than trying to map older county delineations to current ones, Horan and Hargis (1995) sought longitudinal comparability by exploiting the fact that many present-day counties are the result of the partitioning of older, larger counties. Their aggregation strategy therefore groups more recent counties that share common "origins" into larger historical counties that can be compared across time.

Our crosswalk has several advantages over the one provided by Horan and Hargis (1995). First, their approach leads to larger county clusters the longer the time horizon of the analysis. As a result, it does not allow researchers to work with actual present-day counties or commuting zones. For example, to conduct an analysis using data from 1900 to 1990, the data for all decades have to be mapped to the large county groups necessary to accommodate the earliest year, i.e., 1900. Second, Horan and Hargis (1995) rely on information on historical county boundary changes documented by the US Census in writing. When the boundaries of a county change repeatedly it is difficult to keep track of which new county receives the largest share of the old one without a systematic comparison of the respective *maps*. In Appendix D, we provide an example of the limitations of this approach. Our method relies instead on the universe of county maps available for the United States to exactly track *all* changes in county boundaries. We also extend the analysis of Horan and Hargis (1995) 50 years backward (from 1840 to 1790) and 20 years forward (from 1990 to 2010).

The structure of this paper is as follows. Section 2 describes and discusses the method to construct crosswalks. Section 3 applies our method to construct a crosswalk for US counties since 1790 to present-day geographies. We also provide a brief overview of the territorial expansion of the United States, the evolution of its county delineations, and the resulting challenges in constructing a crosswalk for counties. Section 4 concludes.
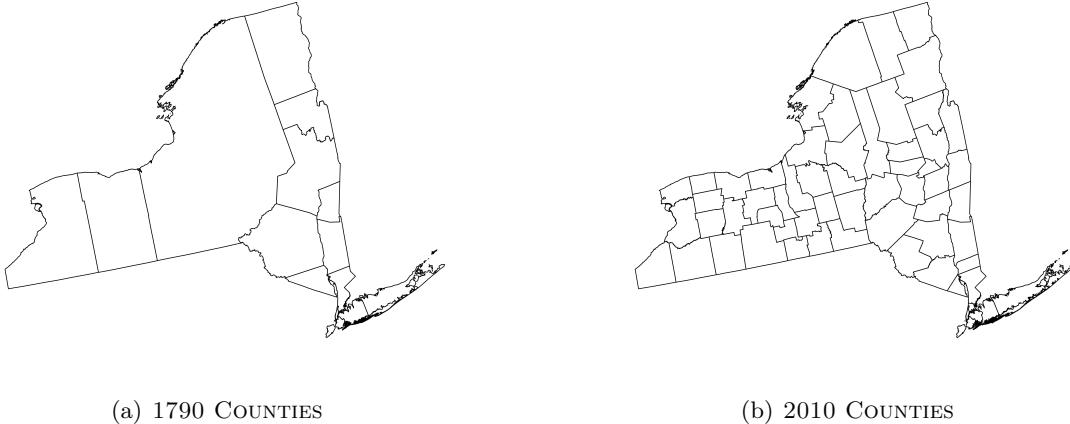
2. A Simple Method to Construct Geographical Crosswalks

In this section, we describe our method to construct geographical crosswalks.

2.1 The Method: Intersecting Different Maps of the Same Territory

Our crosswalk is the result of a simple accounting exercise. We intersect two different maps covering the same territory but partitioned in different ways. As an example, consider the two maps of New York State in Figure 1. The figure displays the county maps for New York in 1790 (left panel) and in 2010 (right panel). Figure 2 shows the intersection of these two maps. The thick gray lines show the 1790 county boundaries, and the thinner perforated lines show the 2010 boundaries.

3

(a) 1790 Counties　　　　　　　　　　　　　　(b) 2010 Counties

Notes: The figure shows the land area of New York State. The left panel shows the county boundaries in New York State in 1790. The right panel shows the county boundaries in New York State in 2010. The underlying shapefiles are taken from Manson et al. (2017).
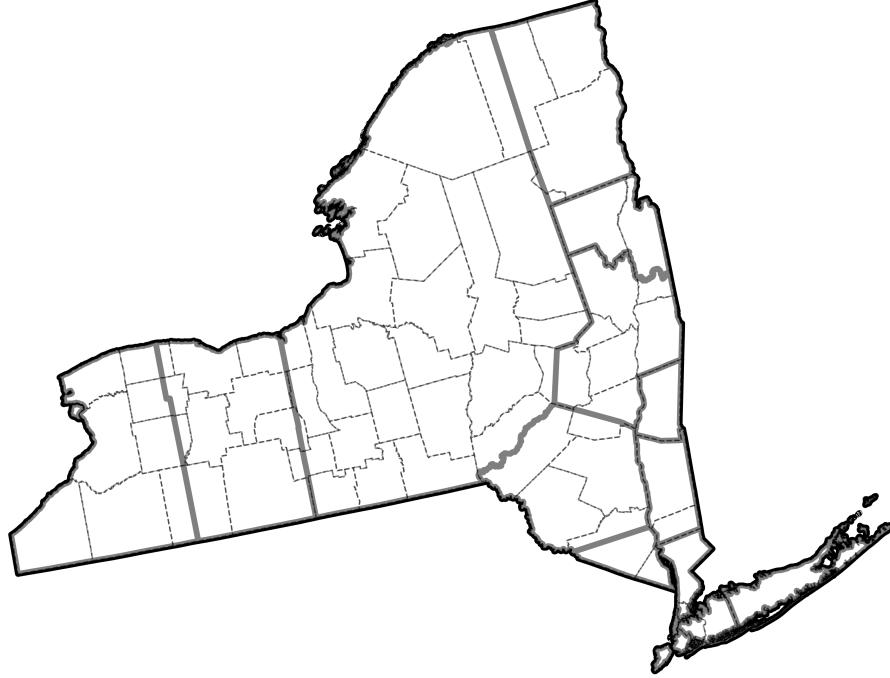
The subunits resulting from the intersection cover the entire land area of New York state and are pairwise disjoint. They can therefore be combined to form any spatial unit in either of the original maps. We use these subunits to construct a collection of weights that constitutes our crosswalk by considering each region in the *reporting* map and computing the share each of its subunits has in its total land area. Each weight is associated with a spatial unit in both the reporting and the reference map and each spatial unit in the two maps has at least one weight associated with it.

To see how to use the resulting crosswalk, consider a specific variable, say population counts by county in 1790. By multiplying the 1790 county population counts for a given county with the associated weights, we uniformly distribute the population across the subunits it contains. We can then sum these population counts across the subunits that are contained within each of the 2010 county delineations. The result is a set of 1790 population counts at the level of 2010 county delineations.

The underlying assumption is that the variable of interest is distributed proportionally to area size across the subunits contained within each spatial unit of the *reporting* map. The plausibility of the assumption depends on the variable of interest. To achieve maximum accuracy, researchers should always use the most disaggregated maps for which data are reported when constructing geographical crosswalks.

We primarily recommend our method to be used with stock data (e.g., data on population or total payroll on the regional level). To map data on average outcomes (e.g., average wages) between different spatial units, a different set of weights is required. They are obtained by considering each

4

Notes: The figure shows the land area of New York State. The gray lines show the county boundaries of 1790. The thinner, perforated black lines show the county boundaries in 2010. The figure shows the spatial subdivisions formed by intersecting the 1790 county map with the 2010 county map. The underlying shapefiles are taken from Manson et al. (2017).

region in the *reference* map and computing the share each of its subunits has in its total land area.[6]

The Python code we provide only produces weights for stock variables since their use is subject to fewer assumptions. The output of the code does, however, report the area of each subunit in the intersected map which researchers can use to compute weights for data on average outcomes. For the rest of the paper, all references to "crosswalk" apply to the collection of weights needed to aggregate stock variables.

Appendix A discusses more details of the crosswalk construction and its use. Appendix B describes how to use the Python implementation of our method to construct a crosswalk given any reporting and reference maps in the shapefile format.

---

[6]In this case we need to assume that the weighting variable underlying the average (e.g., workers in the case of average wages) is distributed proportionally to area size across the subunits contained within each spatial unit of the *reference* map *and* that the averages are constant across the spatial units in the *reporting* map. Note that when the weighting variable is observed, it can always be used to construct stocks (e.g., total payroll by combining total workers with average wages) and our crosswalk can be used as is. We recommend this approach.

Panel A: The Crosswalk for 1810

Year	1810		2010		Weight
	State	County	State	County	
1810	1	A	1	A	1
1810	1	B	1	B	1
1810	1	C	1	C	1

Panel B: The Crosswalk for 1800

Year	1800		2010		Weight
	State	County	State	County	
1800	1	A	1	A	1
1800	1	B	1	A	0.1
1800	1	B	1	B	0.9
1800	1	C	1	C	1

Notes: The table contains an example of the construction of our crosswalk. Panel A contains a hypothetical example of 1810, where the historical counties map directly into the current county borders. Panel B contains an example of 1800, where the original county B was split between the current counties A and B.

2.2 Using the Crosswalk

To illustrate our crosswalk, consider the following hypothetical example. Consider a map of counties in 1800, one in 1810, and a last one in 2010. The goal is to create a crosswalk mapping data provided for the 1800 and 1810 counties to the 2010 county delineations.

Suppose county A took over 10% of county B's land in 1810 and that, since then, there were no further changes. In a first step, we intersect the 2010 county map with the 1810 county map. We notice that the county boundaries overlap perfectly and assign a weight of 1 to each county in 1810, implying the data in 1810 need not be adjusted. See Panel A of Table 1 for the resulting crosswalk for these two counties.

Next, we intersect the 2010 map with the 1800 map. We notice that the 2010 county A consists of the 1800 county A and 10% of the 1800 county B. To account for this, we add a second line for county B as shown in Panel B of Table 1. The first line with a weight of 0.1 indicates the share of the 1800 county B mapping to the 2010 county A; the second line with a weight of 0.9 indicates the share of the 1800 county B that maps to county B in 2010. To use the crosswalk, researchers

should allocate 10% of the population recorded for county B in 1800 to county A in 2010, and 90% to county B.

2.3 Implementing the Method in Python

We implement our method in Python using the "GeoPandas" package.[7] Our program calculates the weights that constitute the crosswalk for two arbitrary shapefiles of the same geography with different partitions (e.g., counties and states of the US). It is essential that these two maps use the same "coordinate system," a property of GIS shapefiles. Users only need to specify which map is the reporting and which the reference map. The code then computes the weights necessary to aggregate data available for the reporting units to the reference units. The online repository associated with this paper (https://www.fpeckert.me/eglp) contains the Python routine.

3. Application: Mapping Historical US County Data to the Present

In this section, we apply our method to construct a crosswalk for US counties since 1790. We first provide a short overview of the history of the US territory and its counties. Then we offer details on the crosswalk construction and discuss caveats.

3.1 The Evolution of the Map of US Counties

Three features of US history are responsible for the vast changes in the geographical delineations of its counties: first, the territorial expansion of the US throughout the 19th century, which increased its territory significantly and led to the creation of *new* counties; second, the conversion of existing US "territories" to "states," which was usually associated with more detailed spatial administrative units; and third, changes in the geographical delineations of the existing county borders, which were adjusted as the population grew. We discuss them in turn.

Territorial Expansion In 1776, the Declaration of Independence turned the thirteen colonies of the United Kingdom on the North American continent into the initial territory of the United States. Between then and now the *territory* of the United States grew gradually into its present shape. Figure 3 depicts the major historical milestones of US territorial history. Following the Declaration of Independence, the purchase of Louisiana from France in 1803 constituted the first major territorial expansion. A set of smaller cessions by Great Britain in 1818 and Spain in 1819 followed. Five major annexations in the 19th century gave rise to the US territory as we know it: the annexation of Texas in 1845, the cession of Oregon territory by Great Britain in 1846,

---

[7]GeoPandas is an open-source project that facilitates work with geospatial data in Python. For the GeoPandas documentation, see http://geopandas.org/.

FIGURE 3: TERRITORIAL HISTORY OF THE UNITED STATES

Notes: The map shows the territorial history of the United States. Source: National Atlas of the United States.
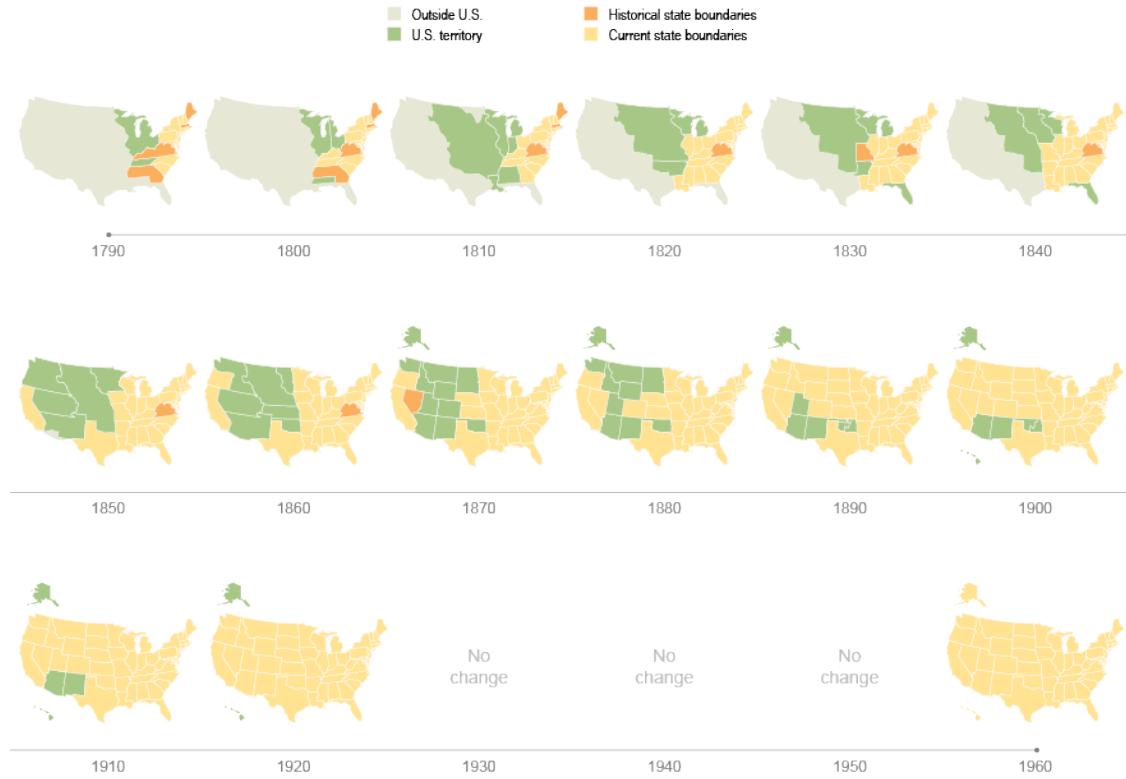
the takeover of Mexican lands in 1848 and 1853, and the purchase of Alaska from Russia in 1867. Lastly, Hawaii was annexed in 1898.

The Creation of the Union Large parts of the United States, however, remained so-called "territories of the United States" even after becoming part of the US proper, i.e., were not yet part of the Union.[8] Figure 4 visualizes the acquisition of new territory and the gradual conversion of these territories into organized states. Each decade before 1920 saw changes in the territorial extent of states. The continental United States *territory* as we know it today was completed in 1900. However, the *Union* of states reached completion much later with the joining of Hawaii in 1959. Table E.2 in Appendix E provides a list of the dates of accession to the Union for each state.

Changing County Boundaries In response to changing social, economic, and political conditions within the US, counties have been created, eliminated, and changed their shape over its history. The partitioning of existing counties to form new ones is the most frequent source of

---

[8]In fact, 31 of the United States' current 50 states were at one time part of such territories. The six states that were not part of the founding states and were never territories are: Kentucky and West Virginia (both split off from Virginia); Maine (split off from Massachusetts); California (created as a state out of the unorganized territory of the Mexican Cession); and Vermont and Texas (both previously self-declared republics).

FIGURE 4: THE TERRITORY AND THE UNION STATES OVER TIME

Notes: This figure shows the evolution of the US territory since 1790. Source: US Census Bureau, Decennial Censuses 1790 to 1960 https://www.census.gov/dataviz/visualizations/048/.

changes in geographical delineations.[9]

Table 2 shows statistics on how the number and shape of counties in the United States have changed throughout its history. The second column lists the number of historical subdivisions in the county shapefiles obtained from Manson et al. (2017). These subdivisions are mostly counties, but also include parishes and independent cities that covered part of the US territory in lieu of counties in various decades. Between 1790 and 2010 the number of such units increased by a factor of more than 10, from 292 to 3143. Their increase is both due to the territorial enlargement of the US and the splitting of existing counties. Column 3 of the table shows the fraction of today's US territory that US states occupied in each decade. In 1790 only 22% of today's territory was part of the United States; the territory was only completed in 1900. Columns 4, 5, and 6 provide a measure for the amount and size of changes in county boundaries in each decade. If two maps perfectly overlap, all weights are 1. If there are primarily small changes, many weights will be very small (0.01, say) and many others very large (0.99, say). In a year with large changes, there will be more

---

[9]See the Atlas of Historical County Boundaries described in Long (1995) for a detailed description of all county boundary changes. The online version of the Atlas can be found at http://publications.newberry.org/ahcbp/index.html.

TABLE 2: CHANGING COUNTY NUMBERS AND BOUNDARIES

Year	Total Subdivisions	Fraction of 2010 Territory	Fraction of Crosswalk Weights *between*		
			0.4 and 0.6	0.2 and 0.8	0.01 and 1
1790	292	0.2221	0.018	0.065	0.265
1800	427	0.232	0.021	0.085	0.291
1810	587	0.4882	0.020	0.071	0.278
1820	779	0.4797	0.021	0.075	0.306
1830	1002	0.4955	0.022	0.084	0.259
1840	1285	0.4945	0.021	0.078	0.210
1850	1632	0.8305	0.015	0.059	0.195
1860	2126	0.8377	0.012	0.042	0.157
1870	2334	0.8377	0.010	0.036	0.129
1880	2614	0.9982	0.010	0.028	0.085
1890	2799	0.9982	0.007	0.020	0.062
1900	2848	1	0.006	0.018	0.050
1910	2963	1	0.003	0.012	0.035
1920	3075	1	0.001	0.003	0.017
1930	3110	1	0	0.001	0.009
1940	3108	1	0	0.001	0.008
1950	3111	1	0	0.001	0.007
1960	3134	1	0	0.001	0.007
1970	3141	1	0	0.001	0.005
1980	3137	1	0	0.001	0.003
1990	3141	1	0	0	0.002
2000	3141	1	0	0	0.002
2010	3143	1	0	0	0

Notes: The table shows statistics describing the territorial expansion of the United Stats throughout its history and the changes in county delineations. The "total subdivisions" count gives the number of counties and other subdivisions (parishes, independent cities) reported in a given decade's Census. We report spatial units available on the entire territory of the United States for which data are available, not just subdivisions of states that are part of the Union in a given year. Column 3 gives the fraction of US territory of 2010 that in a given decade already formed part of the United States. We also include non-incorporated states in the area calculation. Columns 4, 5, and 6 give the fraction of non-unitary weights in the county crosswalk that maps a decade's county delineations to the 2010 ones. For two identical maps all weights are 1. The larger a fraction in the last three columns, the larger the number of differences in county delineations between the counties of that year relative to the county map of 2010.

weights in the intermediate range, e.g., between 0.2 and 0.8. Columns 4, 5, and 6 show the fraction of weights that lie in such intermediate ranges.

3.2 Constructing a County Crosswalk

Our county crosswalk is based on publicly available digital maps of US counties (i.e., shapefiles) for each decade from 1790 to 2010. We downloaded these shapefiles from the NHGIS website (see Manson et al. (2017)).[10] In 2010, we use the 2010 "Topologically Integrated Geographical Encoding and Referencing" (TIGER) line shapefile, created by the Census Bureau. In all other years, we use the 2000 TIGER line shapefiles.[11] For each decade, the shapefiles cover all states in the Union; in addition, they often contain counties in US territories outside of all states in the Union in that year.[12]

We apply the methodology outlined in Section 2 in the following way: for each decade from 1790 to 2000, we intersect the corresponding county shapefile with the 2010 county shapefile. For each county in each decade, we then compute the shares of its area that map to the various counties in the 2010 shapefile. For each historical county, our crosswalk contains these shares, or weights, and lists the corresponding 2010 county. Since many historical counties map into several 2010 counties, the crosswalk often contains many more entries than there are historical or present-day counties. In addition, we repeat the exercise using the maps of 2010 counties and of 1990 commuting zones (see Tolbert and Sizer (1996)) as reference maps. The collection of weights for each decade and each reference map (1990 counties, 2010 counties, 1990 commuting zones) is available in the online repository associated with this paper (https://www.fpeckert.me/eglp).

Column 2 of Table 3 shows the number of 1990 counties that can be mapped to the respective territory of the United States in any given year. Since this holds all geographical boundaries fixed, this provides a good measure of the gradual expansion of the Union of the United States. The largest increase in the number of 1990 counties occurs between 1790 and 1840, when the territory of the US grew the most (cf. Table 2 above). Likewise, once the territory of the US reached its completion in 1900, the number of constant boundary counties remains unchanged. Columns 3 and 4 repeat the exercise for 2010 county delineations and 1990 commuting zones, respectively.

Appendix C provides an overview of the variables in the crosswalk files we provide online and how to use them. Appendix D compares our county crosswalk to the one by Horan and Hargis (1995).

---
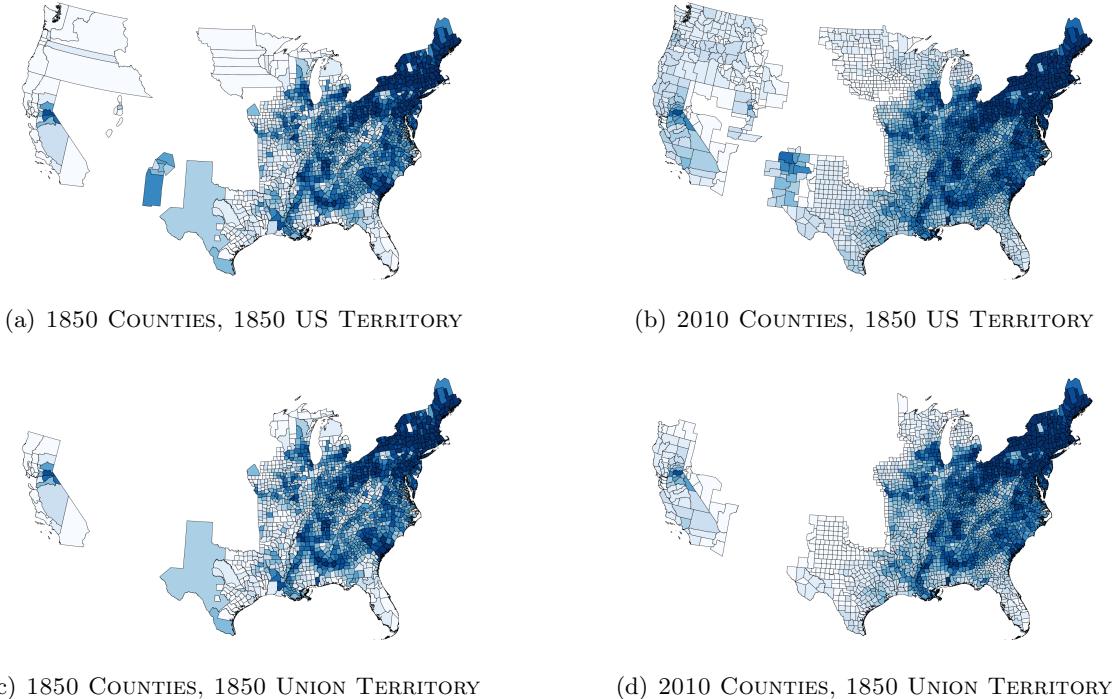
[10]The NHGIS data repository is updated every year and in future versions some details of the shapefiles may change. However, NHGIS assigns a Digital Object Identifier (DOI) code to each version of its data and archives it. The version history for NHGIS is found at https://ipums.org/projects/ipums-nhgis. We used the 2019 version in our work which has been assigned the following DOI code: 10.18128/D050.V14.0.

[11]For 2010 no 2000 TIGER line shapefile exists. For more details on the TIGER system, see Marx (1986).

[12]The counties in these territories tend to be large and often change significantly upon accession to the Union. Arkansas, which joined the Union in 1836, serves as a fitting example. In the 1840 shapefile, Arkansas contains 41 counties. A decade earlier, i.e., in 1830 the Arkansas "territory" contained only 24 counties.

FIGURE 5: COUNTY-LEVEL POPULATION IN 1850
FOR DIFFERENT COUNTY DELINEATIONS

(a) 1850 COUNTIES, 1850 US TERRITORY

(b) 2010 COUNTIES, 1850 US TERRITORY

(c) 1850 COUNTIES, 1850 UNION TERRITORY

(d) 2010 COUNTIES, 1850 UNION TERRITORY

Notes: The figure shows the 1850 population of the United States across various county delineations. Counties in a darker shade of blue contain more population. The top left panel shows the 1850 population across all territories of the United States at the time for the county delineations at the time. The top right panel shows the 1850 population across all territories of the United States at the time for the county delineations of 2010, where the mapping was done using our crosswalk. The bottom left panel shows the 1850 population across only the states in the Union at the time for the county delineations at the time. The bottom right panel shows the 1850 population across only the states in the Union at the time for the county delineations of 2010, where the mapping was done using our crosswalk.

3.3 Example: Mapping Historical US County Populations to Modern-Day Counties and Commuting Zones

In this subsection, we use the county crosswalk to create a longitudinal panel of population counts from 1790 to 2010 on the 2010 county level. Both the data and code can be found in the online repository associated with this paper (https://www.fpeckert.me/eglp).

We use data on historical population counts at the county-level for every decade from 1790 to 2010 obtained from the NHGIS database. Once the data have been merged to the crosswalk, we aggregate population counts across the spatial subunits within each 2010 county for each decade from 1790 to 2010.

The procedure produces a longitudinal, unbalanced panel of US county populations between 1790 and 2010 at the 2010 county level. We calculate the total population we can assign to 2010 counties

TABLE 3: COUNTIES AND COMMUTING
ZONES THROUGHOUT US HISTORY

Year	1990 Counties	2010 Counties	1990 Commuting Zones
1790	1515	1567	306
1800	1562	1622	318
1810	2529	2559	579
1820	2482	2524	566
1830	2549	2577	579
1840	2541	2571	579
1850	3109	3108	722
1860	3111	3109	722
1870	3111	3109	722
1880	3136	3138	737
1890	3136	3138	737
1900	3141	3143	741
1910	3141	3143	741
1920	3141	3143	741
1930	3141	3143	741
1940	3141	3143	741
1950	3141	3143	741
1960	3141	3143	741
1970	3141	3143	741
1980	3141	3143	741
1990	3141	3143	741
2000	3141	3143	741
2010	3141	3143	741

Notes: The table shows the number of contemporary spatial units our crosswalk maps to the territory of US states in various decades. The table hence provides an overview of how many spatial units can consistently be studied from each decade to the present. Column 2 shows the number of 1990 county delineations that can be mapped to the US territory in each decade. Columns 3 and 4 do the same for 2010 counties and commuting zones (see Tolbert and Sizer (1996)).

in each decade and compare these "mapped" population counts to aggregate population counts for the US obtained from US Census publications. In the majority of decades, we can assign the entire US population to 2010 counties. The largest discrepancy occurs in 1790, where we can only map 99.1% of the US population. Details are provided in Table E.1 in Appendix E.

Figure 5 shows the population data for 1850 on a series of different maps. The top left panel shows the 1850 population across all territories of the United States at the level of 1850 county delineations. The top right panel shows the 1850 population across all territories of the United States at the time for the county delineations of 2010, where the mapping was done using our crosswalk. The bottom left panel shows the 1850 population across only the states in the Union at the time for the contemporary county delineations. The bottom right panel shows the 1850 population across only the states in the Union at the time at the level of the county delineations of 2010, where the mapping was again done using our crosswalk.

4. Concluding Remarks

In this paper, we propose and implement a simple method to construct crosswalks between a reporting and a reference map. The method relies on intersecting two maps to create a new map consisting of smaller spatial subunits. For each subunit, we compute the share it accounts for in the overall land area of the spatial unit that nest it in the reporting map. The collection of these shares, or "weights," constitutes a crosswalk from the reporting to the reference map. The weights can be used to re-aggregate data on stocks (e.g., population counts) available for the reporting units to the reference units.

We apply the method to US counties since 1790. We construct a simple-to-use crosswalk that researchers can employ to map historical county-level data for the United States to present-day county delineations or commuting zones. We then apply the crosswalk to historical Census data to create a longitudinal county-level panel of population based on the 2010 county map.

A data repository associated with this project at https://www.fpeckert.me/eglp contains the following material: (i) Python code that allows researchers to take two maps of the same territory and construct their own crosswalk (see Appendix B for a step-by-step instructions), (ii) Python code and crosswalks that map historical US counties since 1790 to 1990 counties, 2010 counties, or 1990 commuting zones and as an example of how to use the county crosswalk, (iii) Python code that maps 1790 county-level population counts to 2010 county delineations.

References

ACEMOGLU, D. AND P. RESTREPO (2017): "Robots and Jobs: Evidence from US Labor Markets," Working Paper 23285, National Bureau of Economic Research.

AUTOR, D. H. AND D. DORN (2013): "The Growth of Low-skill Service Jobs and the Polarization of the US Labor Market," *American Economic Review*, 103, 1553–97.

AUTOR, D. H., D. DORN, AND G. H. HANSON (2013): "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, 103, 2121–68.

ECKERT, F. AND M. PETERS (2018): "Spatial Structural Change," Working Paper.

FORSTALL, R. L. (1996): "Population of States and Counties of the United States: 1790 to 1990," *Deparment of Commerce, U.S. Bureau of the Census Population Division*.

HORAN, P. M. AND P. G. HARGIS (1995): "County Longitudinal Template, 1840-1990." *Ann Arbor, MI: Inter-university Consortium for Political and Social Research*.

LONG, J. H. (1995): "Atlas of Historical County Boundaries," *The Journal of American History*, 81, 1859–1863.

MANSON, S., J. SCHROEDER, D. VAN RIPER, AND S. RUGGLES (2017): "IPUMS National Historical Geographic Information System: Version 12.0," *Minneapolis: University of Minnesota. 2017. http://doi.org/10.18128/D050.V12.0.*

MARX, R. W. (1986): "The TIGER System: Automating the Geographic Structure of the United States Census," *Government Publications Review*, 13, 181–201.

TOLBERT, C. M. AND M. SIZER (1996): "US Commuting Zones and Labor Market Areas: A 1990 Update," *ERS Staff Paper*.

Appendix

A. Details of County Crosswalk Construction

In this subsection, we discuss some details of the crosswalk construction.

FIPS and ICP County Codes Historical data downloaded from NHGIS or other sources often contain several sets of county and state identifiers. The two main classification schemes are the so-called "FIPS" and "ICP" codes. The FIPS codes are the standard codes used in most modern data sets. ICP codes are a coding scheme developed by the Inter-University Consortium for Political and Social Research (ICPSR). These codes, in contrast to FIPS codes, also identify areas that are not part of any county, including the independent cities of Virginia and Indian territories. In addition, they have specific variable codes for missing, edited, or unidentified observations. As a result ICP codes are useful when working with data for counties that may not exist any more today. For *counties* that exist until today, FIPS and ICP county codes coincide. FIPS and ICP *state* codes, however, always differ. Our crosswalk allows researchers to merge data from any decade using either ICP or FIPS codes.

Shapefile Imprecisions We use the Python GeoPandas package to overlay shapefiles for different decades and form their intersection. However, some of the shapefiles contain imprecisions. Sometimes, even county boundaries that have not changed between two decades do not exactly overlap. In other words, the geographical delineations of some unchanged counties differ slightly between two generations of shapefiles. In such cases, our procedure identifies very small weights that would be added as additional entries in the crosswalk. In practice, it is difficult to determine which intersections identify actual county boundary changes and which are just due to such imprecisions. To be conservative, we drop intersections that create areas of less than ten square meters in the construction of our crosswalk. The crosswalk *weights* associated with these dropped areas are at the order of magnitude 10^{-10}, i.e., account for a very small share of the land areas of reporting counties.

Counties, Parishes, and Independent Cities In earlier years, counties did not cover the entire territory of the US. Instead, some parts were covered by parishes, independent cities, and other subdivisions. In fact, some independent cities continue to exist to this day.[13] The crosswalk maps these subdivisions to 2010 counties or independent cities. All these subdivisions are assigned "county" codes under the ICP classification system. Only counties and independent cities that still

---

[13]Formally, an independent city is not part of the territory of any county. There are 41 independent US cities, 38 of which are in Virginia. The three independent cities outside Virginia are Baltimore (Maryland), St. Louis (Missouri), and Carson City (Nevada). The US Census Bureau uses counties as its base unit for presentation of statistical information, and treats independent cities as county equivalents for those purposes.

exist in 2010 are assigned "county" identifiers in 2010. As a result, researchers who are interested in mapping data for such non-county subdivisions should always use the ICP identifiers in the crosswalk.

B. How to Create Customized Crosswalks

Here we describe how to use our programs to create geographical crosswalks between a reporting and reference map of the same territory.

- Go to https://www.fpeckert.me/eglp and download the Python code for creating general geographical crosswalks.

- Ensure that the reporting and reference shapefiles are governed by the same coordinate system and cover the same territory (e.g., the continental US) but contain different partitions of it (e.g., counties and school districts). Note that each shapefile must contain a column that is the unique identifier of each region, e.g., state and county codes in the case of our county crosswalk.

- Copy the *file paths* to the reporting and reference shapefiles into the indicated locations in the Python file (i.e., fill in values for the placeholders "PATH_REPORTING", "PATH_REFERENCE").

- Copy the *file names* of the shapefiles into the indicated locations in the Python file (i.e., fill in values for the placeholders "REPORTING_FNAME", "REFERENCE_FNAME").

- Copy the *column names* of the unique identifier in the two shapefiles into the indicated locations in the Python file (i.e., fill in values for the placeholders "REPORTING_GEOID" and "REFERENCE_GEOID").

- Run the code. For detailed shapefiles (e.g., counties instead of states) the procedure is time-consuming.

- The output produced by the program is a "comma-separated value" file with three columns: (1) the reporting unique identifier, (2) the reference unique identifier, and (3) the associated weights.

C. Description of Crosswalk Files

On the paper's website https://www.fpeckert.me/eglp, we provide three different versions of the crosswalk that map county delineations from 1790 to 2010 to (1) 1990 county delineations, (2) 2010 county delineations, and (3) 1990 commuting zones delineations (Tolbert and Sizer (1996)). Each crosswalk comes in the form of a comma-separated value file that contains the geographical identifiers for each decade from 1790 to 2010, the geographical identifiers of the reference map, and a set of weights. Tables C.1, C.2, and C.3 provide an overview of the variables available in each crosswalk.

To use the crosswalk, researchers first need to determine the decade closest to the data that is to be mapped and choose the appropriate reporting year in the crosswalk. Next, they need to check whether the state and county identifiers in their data are ICP or FIPS codes. Lastly, the user needs to decide which geographical delineation to map the data to: 1990 counties, 2010 counties, or 1990 commuting zones. Once county codes in the crosswalk and the data have been assigned the same name, the user should merge the crosswalk file onto the historical data using a one-to-many merge. The weights should then be applied to the historical data points (e.g., population stocks at the county level); in particular, the variable of interest in the historical data should be multiplied by the weights. The result is a set of population counts for each spatial subunit resulting from the intersection of the reporting map and reference map. These subunit populations can then be summed within the reference counties to yield county-level population counts at the reference county level. Doing this for each decade of historical data produces a longitudinal panel of county-level population counts at 2010 county delineations.

The crosswalk can also be used to concord data on average outcomes rather than stocks (e.g., average wages rather than total payroll). To this end, we include the "area" variable in each crosswalk. It contains the area of the overlap between the reporting and the reference county. For data on averages, the correct weight is the share of the area of the reference county accounted for by the reporting county. Having computed these weights, researchers can re-weight averages for historical counties to obtain them at the level of the reference counties.

The paper's website offers an example of how to use the crosswalk with historical population data from Manson et al. (2017).

TABLE C.1: VARIABLES IN THE 1990 CROSSWALK

Variable	Description
Year	Decade of the historical data; takes values 1790-2010
NHGISST	NHGIS state code of historical county
NHGISCTY	NHGIS county code of historical county
STATENAM	Name of state of historical county
NHGISNAM	Name of state of historical county
ICPSRST	ICPSR state code of historical county
ICPSRCTY	ICPSR county code of historical county
area_base	Area, in square meters, of the historical county
NHGISST_1990	FIPS state code of historical county
NHGISCTY_1990	FIPS county code of reference county
STATENAM_1990	Name of state of reference county
NHGISNAM_1990	Name of state of reference county
ICPSRST_1990	ICPSR state code of reference county
ICPSRCTY_1990	ICPSR county code of reference county
area	Area, in square meters, of the intersection between the historical county and the reference county
weight	Share of historical county's area that overlaps with reference county in 1990
US_STATE	Indicator for whether the US state containing the historical county was part of the Union in the year

Notes: The table shows the variables in the crosswalk that maps US counties from 1790 to 2010 to 1990 county delineations (crosswalk available at https://www.fpeckert.me/eglp).

TABLE C.2: VARIABLES IN THE 2010 COUNTY CROSSWALK

Variable	Description
Year	Decade of the historical data; takes values 1790-2010
NHGISST	NHGIS state code of historical county
NHGISCTY	NHGIS county code of historical county
STATENAM	Name of state of historical county
NHGISNAM	Name of state of historical county
ICPSRST	ICPSR state code of historical county
ICPSRCTY	ICPSR county code of historical county
area_base	Area, in square meters, of the historical county
NHGISST_2010	FIPS state code of historical county
NHGISCTY_2010	FIPS county code of reference county
STATENAM_2010	Name of state of reference county
NHGISNAM_2010	Name of state of reference county
ICPSRST_2010	ICPSR state code of reference county
ICPSRCTY_2010	ICPSR county code of reference county
area	Area, in square meters, of the intersection between the historical county and the reference county
weight	Share of historical county's area that overlaps with reference county in 2010
US_STATE	Indicator for whether the US state containing the historical county was part of the Union in the year

Notes: The table shows the variables in the crosswalk that maps US counties from 1790 to 2010 to 2010 county delineations (crosswalk available at https://www.fpeckert.me/eglp).

Table C.3: Variables in the 1990 Commuting Zone Crosswalk

Variable	Description
Year	Decade of the historical data. Takes values 1790-2010.
NHGISST	FIPS state code of historical county.
NHGISCTY	FIPS county code of historical county.
STATENAM	Name of state of historical county
NHGISNAM	Name of state of historical county
ICPSRST	ICPSR state code of historical county.
ICPSRCTY	ICPSR county code of historical county.
area_base	Area, in square meters, of the historical county
cz	1990 Commuting zone code for reference commuting zone (see Tolbert and Sizer (1996))
area	Area, in square meters, of the intersection between the historical county and the reference county
weight	Share of historical county's area that overlaps with reference commuting zone in 1990.
US_STATE	Indicator for whether the US state containing the historical county was part of the Union in the year

Notes: The table shows the variables in the crosswalk that maps US counties from 1790 to 2010 to 1990 commuting zone delineations as defined by Tolbert and Sizer (1996) (crosswalk available at https://www.fpeckert.me/eglp).

D. An Alternative County Crosswalk

Horan and Hargis (1995) developed the County Longitudinal Template (CLT) for the years from 1840 to 1990. The CLT traces the historical origins of counties to create county groups that can be studied over time. It exploits the fact that many of today's counties resulted from splits of historically larger counties to aggregate data for today's counties to the level of these historical antecedents. In this case no weights are needed; count variables can just be added up across the constituent counties of historically larger ones.

TABLE D.1: EXCERPT FROM THE COUNTY
LONGITUDINAL TEMPLATE:
MONTGOMERY COUNTY (AL) IN 1840

	1840		1990	
Year	State	County	State	County
1840	AL	Montgomery	AL	Autauga
1840	AL	Montgomery	AL	Barbour
1840	AL	Montgomery	AL	Bullock
1840	AL	Montgomery	AL	Butler

Notes: The table shows an excerpt from the County Longitudinal Template by Horan and Hargis (1995). It shows the 1840 Montgomery County and how to allocate its employment to 1990 counties in Alabama. The underlying shapefiles are taken from Manson et al. (2017).

Table D.1 shows an excerpt from the CLT crosswalk. The CLT methodology takes a group of modern counties, such as Autauga, Barbour, Bullock, and Butler counties (all in Alabama), and associates them with a historical county, in this case Montgomery, Alabama, that they used to be part of based on documented boundary changes. Figure D.1 shows the historical Montgomery County in the middle and all the present-day counties that are mapped to it in the CLT crosswalk (Table D.1 only shows a few).
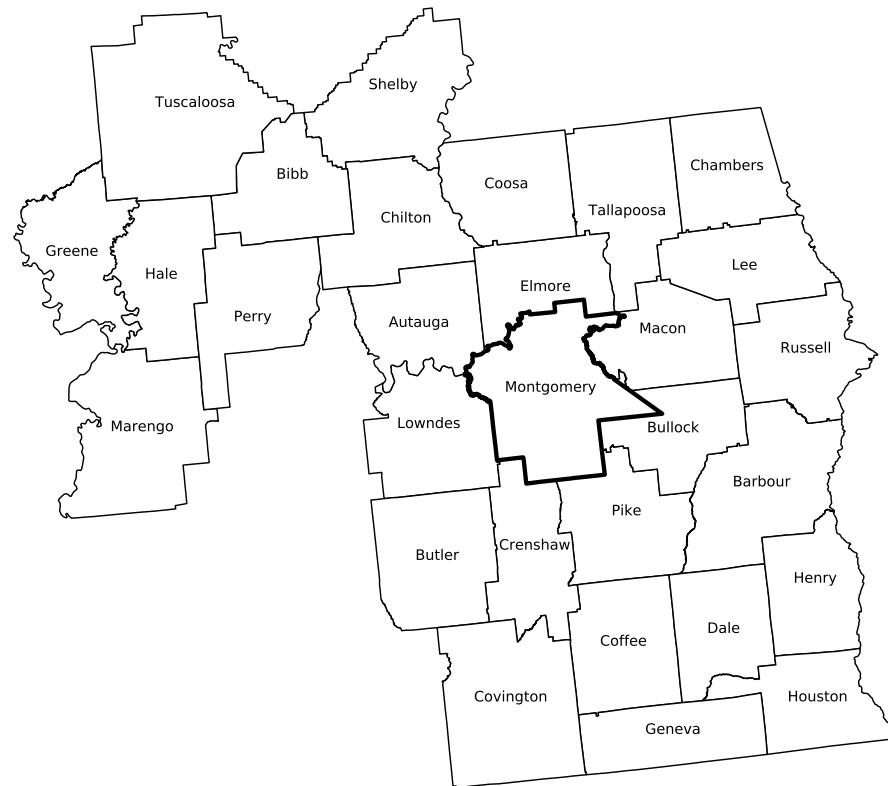
The figure highlights the problems that can occur when using the CLT crosswalk. While the methodology considers the set of all actual documented changes, it does not consider the geographical details of these changes. This results in cases where modern counties such as Barbour, Butler, and Autauga that do not share any common area with the historical county Montgomery are assigned to it nevertheless. This is likely the result of a long series of changes between 1840 and 1990 in which counties were split and boundaries moved many times.

When using our crosswalk to map 1840 county data to 1990 county delineations, historical data for Montgomery county is not assigned to any of Autauga, Barbour, or Butler county. Instead, in our crosswalk, modern Montgomery county is assigned approximately 0.8 of count variables reported for historical Montgomery, reflecting the fact that historical Montgomery county was

situated approximately where modern Montgomery was in 1990. The remaining mass is assigned in almost equal proportions to Bullock and Elmore counties.

Our crosswalk has a range of additional advantages. We extend the crosswalk back 50 years to 1790 and forward 20 years to 2010. Our crosswalk can be used to work with present-day counties and does not necessitate working with historical county groups. With the CLT, the further back in time researchers go the more data for *all* years have to be aggregated, since data always have to be summed to the larger historical counties. With our crosswalk, all data for all decades can be mapped to present-day county boundaries. Furthermore, our crosswalk can be used *both* to aggregate data reported for present-day counties to the delineations of earlier years, and to disaggregate data for historical counties to modern county or commuting zone delineations.

FIGURE D.1: 1840 MONTGOMERY COUNTY AND 1990 COUNTIES IN ALABAMA

Notes: The figure shows the 1990 counties that the Horan and Hargis (1995) crosswalk maps to the area of Montgomery county in 1840. The figure also shows the geographical delineations of Montgomery county in 1840. The underlying shapefiles are taken from Manson et al. (2017).

E. Additional Tables

TABLE E.1: THE AGGREGATE POPULATION IN THE CENSUS
AND FROM NHGIS

Year	Census	NHGIS	Discrepancy	
			Absolute	Percentage (%)
1790	3929214	3893934	35280	0.9
1800	5308483	5299313	9170	0.2
1810	7239881	7238667	1214	0.0
1820	9638453	9621616	16837	0.2
1830	12866020	12860492	5528	0.0
1840	17069453	17063353	6100	0.0
1850	23191876	23191876	0	0.0
1860	31443321	31362622	80699	0.3
1870	38558371	38557210	1161	0.0
1880	50189209	50011451	177758	0.4
1890	62979766	62654302	325464	0.5
1900	76212168	75902898	309270	0.4
1910	92228496	92164175	64321	0.1
1920	106021537	106021568	-31	0.0
1930	122775046	123202660	-427614	-0.3
1940	132164569	132165129	-560	0.0
1950	150697361	151325798	-628437	-0.4
1960	179323175	179323175	0	0.0
1970	203392031	203208441	183590	0.1
1980	226545805	226545805	0	0.0
1990	248709873	248709873	0	0.0
2000	281421906	281421906	0	0.0
2010	308745538	308745538	0	0.0

Notes: The table reports total US population in each decade and the total population that can be assigned to a county using NHGIS data. Columns 4 and 5 also show the discrepancy between aggregate population numbers and what can be attributed to a county using the NHGIS data. The NHGIS data are taken from the NHGIS website (see Manson et al. (2017)). The Census data for 1790-1990 are taken from Forstall (1996).

TABLE E.2: ACCESSION OF US STATES TO THE UNION

State	Union Accession	State	Union Accession
Delaware	1787	Michigan	1837
Pennsylvania	1787	Florida	1845
New Jersey	1787	Texas	1845
Georgia	1788	Iowa	1846
Connecticut	1788	Wisconsin	1848
Massachusetts	1788	California	1850
Maryland	1788	Minnesota	1858
South Carolina	1788	Oregon	1859
New Hampshire	1788	Kansas	1861
Virginia	1788	West Virginia	1863
New York	1788	Nevada	1864
North Carolina	1789	Nebraska	1867
Rhode Island	1790	Colorado	1876
Vermont	1791	North Dakota	1889
Kentucky	1792	South Dakota	1889
Tennessee	1796	Montana	1889
Ohio	1803	Washington	1889
Lousiana	1812	Idaho	1890
Indiana	1816	Wyoming	1890
Mississippi	1817	Utah	1896
Illinois	1818	Oklahoma	1907
Alabama	1819	New Mexico	1912
Maine	1820	Arizona	1912
Missouri	1821	Alaska	1959
Arkansas	1836	Hawaii	1959

Notes: The table reports the year in which individual states became part of the Union of the United States, i.e., became US states as we know them today. States are ordered by their date of accession to the Union.