A Consistent County-Level Crosswalk for US Spatial Data since 1790*

Fabian Eckert Yale University Andrés Gvirtz Yale University Michael Peters Yale University

July 2018

Abstract

Both the number and the geographic boundaries of the counties covering the territory of the United States changed frequently and significantly since 1790. In order to study the spatial aspects of the United States' economic history it is frequently necessary to create a consistent panel of spatial units, which are consistent over time. We provide such a crosswalk, that enables researchers to aggregate historical US county data for every decade since 1790 to current US counties and commuting zones. This note describes the details of our data construction and how to use this crosswalk in practice. A data package containing the decadal crosswalks since 1790 and the accompanying GIS files is available on our websites.

^{*}Corresponding author: fabian.eckert@yale.edu. We thank the Yale StatLab and in particular Henry Glick for help with ArcGIS. We are also grateful to Dave van Riper from the Minnesota Population Center for numerous discussions. All errors are our own.

1 Introduction

An increasing number of research in the social sciences is concerned with spatial aspects of the historical development of the United States. A frequent issue in working with such detailed historical geographic data are changes in the geographic boundaries of the spatial units of interest.

The dominant source for historical geographical data of the United States is the National Historical Geographic Information System (NHGIS) website.¹ While NHGIS provides a vast amount of historical information at the spatial level, it is not straightforward to combine this information in a panel format. The reason is that NHGIS provides data with geographical identifiers at the time of data *collection*. These spatial units are *not* consistent over time given the enormous changes in most spatial delineations throughout US history. These changes are both due to the territorial expansion of the United States, which led to the creation of more and more administrative spatial units, and to its population growth, which regularly caused existing administrative spatial units to split so as to maintain a similar population distribution across administrative entities. In order to study the historical evolution of particular spatial units, one therefore requires a crosswalk, which takes such border changes into account. In this note and the accompanying dataset, we provide such a crosswalk.

We focus on one particular administrative spatial unit: counties. The reason is that US counties are the smallest spatial unit for which historical information is available in a consistent manner. The number of counties rose from about 250 in 1790, spread over the initial 13 founding states, to more than 3000, covering the United States territory in 2000. In this note, we describe the construction and use of a data crosswalk from historical county boundaries in every decade from 1790-2000 to the 1990 county borders. The crosswalk and additional materials are made freely available to researchers and can be downloaded online from our websites. Given the recent interest in regional labor markets in the economics literature we also include a crosswalk from US counties as of 1990 to commuting zones as described in Tolbert and Sizer (1996). The crosswalk will hence allow researchers to aggregate spatial historical data on the United States to spatial units that are constant throughout the history of the United States from 1790 to 2000, allowing them to meaningfully analyze the evolution of the spatial economy of the United States through time and space.

We are not the first to construct such a crosswalk. In 1995, Horan and Hargis (1995) published

¹In its own words it "provides population, housing, agricultural, and economic data, along with GIS-compatible boundary files, for geographic units in the United States from 1790 to the present".

a County Longitudinal template in order to allow for an inter-temporal comparison of fixed county groups between 1840-1990. Their approach pre-dates the use of ARCGIS software. Instead, the authors tried to ensure longitudinal comparability through an aggregation strategy which groups counties, as defined by their boundaries in 1990, into larger units on the basis of earlier historical county boundary configurations. Naturally, these "county groups/clusters" become larger the further the researcher intends to go back in time. In some cases, the scale of the historically comparable area becomes unwieldy. In particular, geographical areas that have witnessed the proliferation of new counties are the most likely to suffer from such aggregation problems.

While Horan and Hargis (1995) aggregate counties which have common historical origins, we provide a crosswalk that disaggregates historical counties into 1990 county boundaries based on their land area. This is achieved by using ArcGIS software to overlay historical county shape files with the 1990 county shape file and then calculating the share of land of each historical county that forms part of a given recent county. Using such land partitions as weights, researchers can directly aggregate historical information to spatial units as defined by their recent boundaries. While the approach of Horan and Hargis (1995) leads to larger county clusters, the longer the time horizon of the analysis, our approach keeps the number of counties constant. As a result, our crosswalk allows researchers to analyze the data from 1790 to 2000 in a consistent way. In addition to the finer spatial granularity, we also extend the analysis by Horan and Hargis (1995) by 50 years. Having mapped the historical data to recent county borders it is then easy to further aggregate the data to present day larger spatial units such as PUMAs, commuting zones or Census regions. We include a mapping to 1990 commuting zones by Tolbert and Sizer (1996) in our crosswalk for convenience and also provide the associated shape files.

The crosswalk described in this note was constructed for the empirical analysis in Eckert and Peters (2018). Eckert and Peters (2018) aggregate historical county level data to 1990 commuting zone level to study the joint movement of US workers across sectoral and spatial lines during the structural transformation, i.e. between 1880 and 2000. The structure of this note is as follows. Section 2 provides a brief overview over the territorial expansion of the United States and the evolution of county delineations. Section 3 describes the construction of the crosswalk and discusses some caveats. Finally, Section 4 describes how to use the crosswalk with historical data. Section 5 offers concluding remarks.

2 The Geography of the United States since 1790

There are two important aspects to the geographical history of the United States that markedly affected the organization of administrative units. First of all, the territorial expansion of the United States throughout the 19th Century increased its territory significantly and lead to the creation of new counties. Moreover, the process by which different territories of the US joined the Union and where subsequently subject to the administrative territorial order imposed by the central government and the US Census Bureau was very gradual. Secondly, even within the boundaries of the United States, the existing county borders were subject to frequent change. In particular, various existing counties were split into several sub-entities due to population growth and other counties were fully or partially merged.

2.1 Territorial History

The Declaration of Independence signed in Philadelphia on July 4, 1776, established the thirteen colonies of the United Kingdom on the North American continent as the initial territory of the United States. Over the subsequent roughly 250 years the territory of the United States grew gradually into what we know today as the United States of America. Figure 1 displays the territorial history of the United States. Following the Declaration of Independence, the purchase of Louisiana from France in 1813 constituted the first major territorial expansion. This was followed by a set of smaller cessions by Great Britain in 1818 and Spain in 1819. Four major annexations in the 19th century finally constituted the territory how we know it today: the annexation of Texas in 1845, the cession of Oregon territory by Great Britain in 1846, the Mexican cessions from 1848/1853 and finally the purchase of Alaska from Russia in 1867. Hawaii was annexed in 1898.



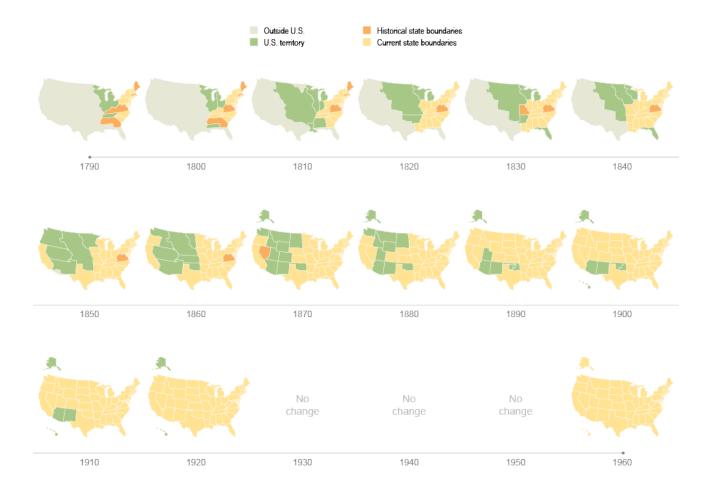
Notes: The figure shows the evolution of the territory of the United States. Source: National Atlas of the United States.

Figure 1: Territorial History of the United States

Large parts of the United States, however, remained the so-called "territories of the United States" even after becoming part of US proper, but not yet part of the Union.² Such territories were then gradually admitted to become states os the US. Usually the most populous territories would be the first to achieve statehood.³ Figure 2 visualizes the acquisition of new territory and the conversion of this land into organized states.

²In fact, 31 of the United States' current 50 states were at one time part of such territories.

³The 8 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).



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

Figure 2: Territorial History of the United States

It is apparent that this happened at a rapid pace. Each decade before 1920 saw changes in the territorial extent or the make-up of states in the country. The year 1920 marks the year that completes the continental United States territory as we know it today. This was followed by several decades during which there were no changes until Alaska and Hawaii were admitted as states in the 1950s. Figure 5 in the Appendix contains a list of states that form part of the Union for each decade, as well as individual states' date of accession.

It appears that territories varied in the sophistication of their spatial organization. The data on the NHGIS website often includes information for some or all territories at a given point in time. In the construction of our crosswalk we therefore differentiate between the case of using either "all data" available in a year or restricting attention to the states that formed part of the Union at any given point in time.

2.2 County History

The main obstacle for a consistent empirical analysis of local areas in a historical longitudinal context is that the geographic configurations of counties changed regularly. In response to changing social, economic and political conditions, all states have altered county boundaries, eliminated existing counties or created new ones. While some changed involved only transfers of small areas among adjacent counties, the most common form of change was the partitioning of existing counties to form new ones.⁴ Column 2 of Table 1 shows the number of state subdivision for which the Census Bureau reported data in each given year. These subdivisions include Parishes, Independent cities and other spatial units. The actual number of historical counties is hence significantly lower. While these numbers show the significant growth in the number of countries before 1890, they hide the extent of reorganization and division of existing counties. Column 3 of Table 1 shows the number of modern counties that can be mapped to the respective territory of the United States in any given year. Because these countries are - by construction - constant over time, this is exactly the number of counties for which a consistent longitudinal dataset can be constructed with our crosswalk.

3 Construction of the Crosswalk

In this section we describe the construction of our crosswalk. In particular, we provide details on the source data (the "data inputs") and the choices we made to construct the crosswalk.

3.1 Data Inputs

Our crosswalk is based on the publicly available county shape files of the United States from 1790 to 2010. These files are available from the NHGIS website (Manson et al., 2017). For convenience we include these shape files in the download for our crosswalk. The files cover - for each decade - all states in the Union. In addition, they often also contain counties in US territories outside of the Union. Hence, our crosswalk contains (for each decade) all counties within the Union, as well as selected counties in outside territories.

⁴See the Atlas of Historical County Boundaries described in Long (1995) for a collection of the totality of all county boundary changes. The online version of the Atlas can be found here: http://publications.newberry.org/ahcbp/index.html.

Year	Historical Subdivisions	1990 Counties	Commuting Zones
1790	292	255	146
1800	419	382	254
1810	574	495	299
1820	759	701	341
1830	988	905	369
1840	1279	1175	394
1850	1623	1528	578
1860	2080	1871	598
1870	2295	2133	683
1880	2570	2377	719
1890	2813	2671	728
1900	2862	2755	741
1910	2962	2896	726
1920	3076	3054	741
1930	3110	3088	741
1940	3108	3088	741
1950	3111	3092	741
1960	3133	3141	741
1970	3142	3141	741
1980	3137	3141	741
1990	3141	3141	741

Notes: The "historical subdivisions" counts given above are the totals of counties and other subdivisions (parishes, independent cities, etc.) reported in each census. They include a few cases where the population given in the table is 0. The count for 1860 includes 1 each for Colorado and South Dakota (Dakota Territory), which were reported as single units in that census rather than by counties. We could not find counts of actual counties, which are a strict subset of these entities. The census state tables include 3279 counties and other subdivisions, comprising the 3141 counties reported in 1990 and 138 reported only in one or more prior censuses. The "1990 Counties" column list the number of 1990 counties the crosswalk described in this paper maps the historical subdivision to, while the "Commuting Zones" column reports the number of commuting zones they map to. These two numbers are lower than their 1990 counterparts until 1920 when the Census started reporting numbers for the entire territory of the United States as we know it today.

Table 1: The number of counties and commuting zones throughout US History

3.2 Overlaying County Shape Files

Our crosswalk is based on geometric intersection, i.e. we overlay the county maps of each individual year with the baseline map of 1990. In essence, this is an accounting exercise, calculating the fraction of land in a given historical county that forms part of any given 1990 county.⁵ An example is used for illustration: Suppose county A incorporates 10% of county B's land in 1809 and that since then their were no further changes in the borders of counties A and B. In a first step, we intersect the 1990 county map with the 1810 county map. We notice that the county boundaries overlap perfectly and assign a weight of 1, meaning that data in 1810 does not need to be adjusted to be compared consistently with 1990 data. See Table 2a for the resulting structure of the crosswalk for these two counties.

Next we intersect the 1990 map with the 1800 map. We notice that the 1990 county A consists of the 1800 county A plus 10% of the 1800 county B, so we add a second line for county B, one with weight 0.9 mapping to 1990 county B and one with weight 0.1 mapping to county A in 1990 (see Table 2b). The available county level data for 1800 can then be aggregated to 1990 county borders using the respective areas as a weight. Note that this procedure implicitly assumes that the variable of interest is uniformly distributed across space. By applying this cross-walk to subsequent cross-section, this procedure generates a longitudinal panel dataset with consistent county-boundaries.

3.3 Discussion

Here we discuss a few additional considerations in the use and construction of the crosswalk.

3.3.1 ICP and FIPS Codes

Historical data downloaded from NHGIS usually contains several county and state identifiers. The two main classification schemes are the so-called ICP and FIPS codes. The FIPS codes are much wider known. ICP codes are a coding scheme developed by the Inter-University Consor-

⁵This approach is more precise and less error prone than name matching both because new counties often have no direct counterpart in the earlier data and because counties can change their geography quite dramatically. It is also possible for the same name to denote totally different physical areas at different points in time. An example, cited in Horan and Hargis (1995), is William County, North Dakota which was annexed to Mercer and Stark Counties in 1899. The name "William County" therefore ceased to exist. However, within the year, in a different section of the state, a "new" William County was established. Consequently, the same county name appears in two successive decennial censuses, but it refers to completely separate geographic areas.

year	fips_state	county_state	fips_state_1990	fips_county_19	90 weight
1810	1	А	1	А	1
1810	1	В	1	В	1
1810	1	С	1	С	1
		(a) Tł	ne Crosswalk in 1810)	
ye	ear statefi	ps countyfips	statefips1990	countyfips1990	weight
	ear <i>statefi</i> 300 1	ps countyfips A	<i>statefips</i> 1990 1	countyfips1990 A	weight 1
18			<i>statefips</i> 1990 1 1		weight 1 0.1
	300 1	A	<i>statefips</i> 1990 1 1 1 1	A	1
18 18 18	300 1 300 1	A B	<i>statefips</i> 1990 1 1 1 1 1	A A	1 0.1

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.

Table 2: Example Structure of the Crosswalk

tium for Political and Social Research (ICPSR). ICP codes, contrary to FIPS codes, also identify areas that were not part of any county, including the independent cities of Virginia and some Indian lands. In addition, they have specific variable codes for missing, edited, or unidentified observations or observations, which are not applicable (N/A). As such ICP codes lend themselves to data reported on counties that may not necessarily exist any more. Where the county existed and keeps existing throughout the data set, FIPS and ICP county codes coincide. Note that ICP and FIPS state codes always differ. Our crosswalk allows researchers to merge data from any decade using either ICP or FIPS codes and allows to collapse data to either ICP or FIPS state and county codes in 1990.

3.3.2 NHGIS data and the Union

As discussed above, the shape files used in the construction of the crosswalk often contain information about US territories that have not yet joined the Union. In our experience, the data for these territories is typically less reliable than the comparable data for Union states. Likewise the counties in these territories tend to be large and change significantly upon accession to the Union.⁶ Nevertheless, we kept all county identifiers in the basic version of the crosswalk to allow researchers to map territorial data to current counties and commuting zones.

⁶Take Arkansas, which joined the Union in 1836, as an example. In the 1840 shape file, Arkansas contains 41 counties. A decade earlier, i.e. in the 1830 shape file, the Arkansas territory has only 24 counties.

3.3.3 Tiny weights in the crosswalk

We use ARCGIS software to overlay yearly shape files and form their intersection. At times even county boundaries that have not changed show slight discrepancies between the respective historical shape files and the shape file for the recent periods. This results in our crosswalk containing some instances, where part of a county that existed prior to 1990 is to be split into a tiny part going to one 1990 county, with the remainder mapping to another. It is impossible for us to determine whether these splits are due to computational problems in ARCGIS or due to inaccuracies of the historical maps (we suspect the latter is the case). We decided to keep these small weights in the crosswalk and leave it to the user to decide whether to apply them or not. Precisely because such weights are tiny, this should not make a difference in practice.

3.3.4 Disappearing and appearing counties

If one uses our crosswalk to produce a panel of counties starting in 1790, one will notice that some counties experience a process of "switching", where they appear to have data in certain years, then they disappear from the data before re-emerging and staying in later cross-sections. This is related to the discussion in section 3.3.3: if the shape file of a given year is not very precise, a currently existing county may in that particular decade lose a minor fraction of its population to a county which in reality is only created much later. To avoid this, we give researchers the option to drop rows in the crosswalk for which weights are miniscule. Again, this should be inconsequential quantitatively.

4 How to use the Crosswalk

In this section we describe how to use our crosswalk to construct county-level data for the United States from 1790 to 2000. All files referenced in this section can be found on the authors' websites.

4.1 Description of Files

We provide a separate STATA data file for each year from 1790-2000. Each file is called "egp_ctyXctycz1990 where X is the year of the data the user wants to map to the 1990 county boundaries. Each such

file contains 11 data columns. The first column is the year of the data that is to be mapped to 1990 county delineations. Columns 2 and 3 contain fips state and county codes and 4 and 5 ICP state and county codes for the year of the data. Columns 6 to 9 contain 1990 FIPS and ICP county and state codes. Column 10 contains commuting zone codes as constructed by Tolbert and Sizer (1996). Column 11 contains the weight variable, which denotes the fraction of the area of a county in the data that maps to a given 1990 county delineation. Note that every 1990 county is assigned to exactly one 1990 commuting zone.⁷ Figures 2a and 2b provide a simplified (omitting some variables) overview over the layout of the crosswalk files.

4.2 Using the Crosswalk

To use the crosswalk, the user first needs to determine the decade closest to the data he/she is using. Next, the user needs to determine whether the state and county codes in his/her data follow the ICP or FIPS convention and rename them as either "fips_state" or "icp_state" and "fips_county" or "icp_county".

Note that in the crosswalk file each combination of county-state codes will appear several times. In particular it will appear as many times as the number of counties in the 1990 cross-section that contain part of this original county. This implies the user needs to use the "one to many merge" option in STATA. Figure 2b illustrates such a case. Here part of the 1800 county B is part of the 1990 county A in year 1810, while another fraction of it forms part of the 1990 county B.

Once the data has been merged to the crosswalk, one only has to calculate the *weighted* sum for each county or commuting zone in 1990, where the weight is given in column 11. This procedure results in a cross-section of data on the 1990 county or commuting zone level. Note that the number of counties or commuting zones mirrors the territorial coverage of the underlying, original data source. Hence, columns 3 and 4 of Table 1 above give an indication of the maximum number of 1990 counties and commuting zones obtainable for each decade.

⁷Tolbert and Sizer (1996) construct commuting zones using 1990 county to county commuting flow data. As a result every county is assigned to exactly one commuting cluster, called commuting zone.

4.3 An Example: The US Population at the County and Commuting Zone Level since 1790

As an application of our crosswalk, we consider the construction of consistent population data at the county and commuting zone level since 1790. The data on historical population counts at the county level for every decade from 1790 to 2000 stems from NHGIS. The data folder contains both the do file and the raw population data. In Figure 3 we depict the population data aggregated to 1990 commuting zones for each decade between 1790 and 1990. In Figure 4 we show the same data, but in each year we drop data for all states and territories, which are not part of the Union in the given year. The detailed time series of the various states' accession to the Union is contained in Figure 5.

Column 3 of Table 3 in the Appendix shows the total US population when we use the maximum amount of data available from NHGIS that we can map to present day counties using ICP state and county classification. Column 2 contains the official aggregate population counts of the United States published by the US Census Bureau. Columns 3 and 4 show the absolute and relative difference between the these two numbers. In most years the discrepancy is minimal, i.e. the population data available from NHGIS, which we can map in a consistent way agrees with the aggregate population data quite closely.⁸

5 Concluding Remarks

We hope that the crosswalk files described in this note will be of use to researchers interested in the spatial dimension of the (economic) history of the United States. All work has been done to the best of our ability, but we cannot guarantee for the accuracy of these files.

⁸There are two main reasons why the the population data provided by NHGIS does not accord with the official Census data. First of all, some counties are simply not part of the historical maps even though they do appear in some population tabulations. This does not happen often and we only found this happening for counties that ceased to exist. Secondly, we drop all counties if the ICP code is not available. We do so, because not having an ICP code is often an indicator for questionable data quality, especially in early years. Cases are where some counties include an entire territory, which is already counted elsewhere.

References

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

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

A Additional Tables and Figures

Year	Census	Data from NHGIS	Discrepancy	
			Absolute	Relative
1790	3929214	3893934	35280	0,9%
1800	5308483	5298905	9578	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	31328977	114344	0,4%
1870	38558371	38557210	1161	0,0%
1880	50189209	50011393	177816	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	203179838	212193	0,1%
1980	226545805	226545805		0,0%
1990	248709873	247936269	773604	0,3%

Notes: The table reports the total population size as reported from the Census (Column 2), as aggregated from the spatial information available from NHGIS (Column 3) and the discrepancy between the two.

Table 3: The Aggregate Population in the Census and from NHGIS



Notes: We depict the total population for each year at the level of 1990 commuting zones. The data stems from NHGIS (see Manson et al. (2017)). The aggregation to 1990 commuting zones has been performed using our crosswalk.

Figure 3: Total population at the level of commuting zones: 1790 - 2000



Notes: We depict total population for each year at the level of 1990 commuting zones. We focus on the territory, which is part of the Union at each point in time. The data stems from NHGIS (see Manson et al. (2017)). The aggregation to 1990 commuting zones has been performed using our crosswalk.

Figure 4: Total population at the level of commuting zones for Union territories: 1790 - 2000

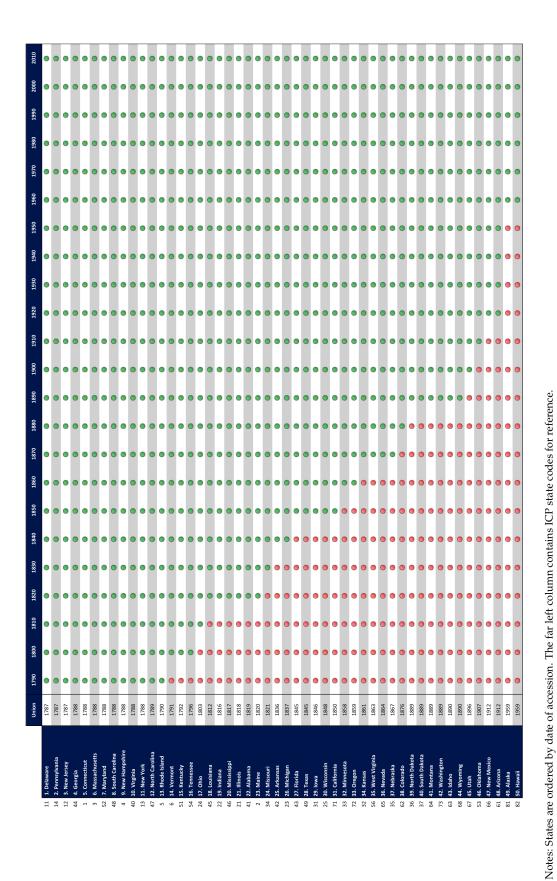


Figure 5: Accession of US States to the Union: 1790-2010