U.S. State-Level Business Cycles Since the Civil War*

Joseph Hoon† Chang Liu‡ Karsten Müller§ Zhongxi Zheng¶

July 6, 2025

Abstract

We construct a novel dataset of 60 macroeconomic time series at the U.S. state level, spanning from the 1863 to the present, based on digitizing and harmonizing 113 historical sources. Equipped with these data, we estimate an annual index of state-level economic activity over nearly 160 years. This index aligns closely with official indicators such as state GDP and unemployment when available. Using this measure of economic activity, we uncover several new facts about state-level business cycles: (1) there is substantial heterogeneity across states in both cyclical dynamics and their underlying drivers; (2) business cycles have become more synchronized since World War II; and (3) downturns have become shorter and recoveries quicker over time.

Keywords: state-level business cycles; economic activity index; mixed-frequency dynamic factor model

JEL classification: C38, E32, N91, N92

*We are grateful to Robert Barro, Arnaud Costinot, Jonathon Hazell, Atif Mian, Emi Nakamura, Jonathan Payne, Christina Romer, Tom Sargent, Jón Steinsson, Bálint Szőke, and Christian Wolf for their invaluable comments and suggestions. We thank Enrico Berkes, Howard Bodenhorn, Sacha Dray, Monty Hindman, Michael McMahon, Varun Sharma, John Wallis, and Thomas Weiss for their assistance with data. We also acknowledge seminar and conference participants at Tsinghua University, Peking University, NUS RMI Macro Workshop, and NUS Macro Brownbag for helpful discussions. Liu acknowledges Singapore MOE AcRF Tier 1 Grant (FY2023-FRC1-004) for financial support. All errors are ours.

†Department of Economics, National University of Singapore. Email: joseph.hoon@u.nus.edu.

‡Department of Economics and Risk Management Institute, National University of Singapore. Email: charlesliu.pku@gmail.com.

§Department of Finance and Risk Management Institute, National University of Singapore. Email: kmueller@nus.edu.sg.

¶Department of Economics, National University of Singapore. Email: zhongxi.zheng@u.nus.edu.

1 Introduction

Reliable indicators on the state of the macroeconomy is the currency of research in empirical macroeconomics, economic history, and growth. Even for an advanced economy like the United States, such data is not always readily available, especially when going back in time. This is even more true for regional data, which has long been of interest to macroeconomists (e.g., Barro and Sala-i-Martin, 1991, 1992; Blanchard and Katz, 1992), and which is increasingly used to identify causal effects (Nakamura and Steinsson, 2018).

The availability of state-level economic data for the United States—particularly for annual data in historical periods—remains limited. For example, the Bureau of Economic Analysis (BEA) publishes annual estimates of state-level GDP only starting from 1963. Before that, there is no state-level annual measure of economic activity, except for a few indicators that capture limited dimensions of it, such as personal income (since 1929), agricultural output (since 1924), and value added of the manufacturing sector (since 1949). As a result, there are many open questions about state-level business cycles and growth: When did which state experience an economic downturn? How do state business cycles differ and to what extent do they coincide with national cycles? And how have state-level business cycles evolved over the long run?

This paper aims to address these questions by constructing a novel dataset containing a variety of state-level economic indicators spanning from 1863 to 2021. Based on an extensive effort to digitize historical publications by U.S. federal and state government agencies and building on the work of other economic historians, we construct a harmonized dataset covering 60 variables. In this dataset, only around 22% of the observations we assemble are available from existing official statistics; the remainder are newly digitized or assembled from various official or private sources. In many cases, we traced the availability of statistics on the production of individual mining products or state government finances through reports published by individual states. We document how we build these time series from 113 sources in a dedicated data appendix that also details the adjustments and imputations required to ensure data consistency. We believe this new dataset has many potential applications in fields such as macroeconomics, development economics, and economic history.

Equipped with our dataset covering over 150 years of U.S. economic history, we estimate an annual index of state-level economic activity covering 1871-2021. To the best of our knowledge, this is the first attempt to estimate state-level economic activity over such a long time. We build on

the existing literature following the spirit of Burns and Mitchell (1946) and view business cycles as common fluctuations in many underlying indicators, which naturally suggests the application of a factor model. In particular, we use a mixed-frequency dynamic factor model similar to Baumeister, Leiva-León and Sims (2024), adapted to our dataset with mixed frequency both within and across variables, to estimate an index from a set of 16 core indicators for each state. For our baseline estimation, these indicators include real activity measures such as output in the agriculture, mining, and manufacturing sectors, as well as data on local labor markets, wealth, government debt and revenues, housing, and transportation.

We confirm the validity of our index by comparing it with existing state-level indicators for the modern period. The index exhibits strong positive correlations with GDP, personal income, unemployment rates, and state coincident indexes. Moreover, our index is also a highly statistically significant and economically meaningful predictor of economic variables *not* used in the construction of the index, such as the number of business failures and bankruptcies, or the number of patents. These findings lend credence to the reliability of the economic activity index in capturing state-level business cycle fluctuations.

Our estimated long-run state-level economic activity index sheds light on the variation of local business cycles across time and space. Three observations stand out from our analysis. First, the structure of state-level business cycles has changed over time, in line with national changes. Before 1950, recessions were longer and recoveries slower, and often concentrated in specific regions. After World War II, economic downturns became shorter and recoveries faster, perhaps they were counteracted by changes in monetary policy, fiscal policy, and broader economic diversification. Second, the co-movement between state-level economic activity and the nationwide business cycle differs considerably across states. This meshes well with existing work by Owyang, Piger and Wall (2005), but we establish it using our long-run time series dating back to the Civil War, which equips us with additional statistical power.

Third, state-level business cycles have become more synchronized, especially since the post-war period. We examine two statistics to track variation in synchronicity over time. We begin by calculating the dispersion of the index across states, which directly measures the extent of variation in economic conditions across states in a given year. As an alternative measure, we follow Kalemli-Özcan, Papaioannou and Peydró (2013) and calculate a synchronization measure for each state as the sum of negative absolute differences between the state's economic activity index and those of all other states in a given year. Intuitively, this measures how each state is different from every

other state. For both measures, we observe large increases in business cycle synchronization across states since World War II.

Given the documented considerable heterogeneity in state business cycles, our state economic activity index lends itself to constructing indicators for state-level recessions by applying existing algorithms such as that of Bry and Boschan (1971*b*). Preliminary findings suggests that, while many recessions align with NBER recession dates, there are also many "forgotten recessions," with more localized downturns. Going forward, this list of recession dates will be potentially useful in the study of local business cycle dynamics among other related applications.

Literature. The primary contribution of this paper is to introduce a novel state-level dataset for the United States comprising dozens of indicators since just after the Civil War, and using these time series to estimate an indicator of regional economic activity. Our work mainly builds upon three strands of literature.

First, we contribute to the literature on historical U.S. business cycle fluctuations. Davis (2004) constructs a measure of U.S. industrial production for 1790-1915, which in turn builds on previous efforts including, among others, Frickey (1947), Romer (1989) and Miron and Romer (1990). While our focus is on constructing regional time series, our work is close to the spirit of this literature in attempting to overcome the limitations of existing data through a large-scale effort to digitize and harmonize information from many sources. Our work is also related to a voluminous literature investigating the properties of the U.S. business cycle (e.g., Long and Plosser, 1983; DeLong and Summers, 1986; Hodrick and Prescott, 1997; Stock and Watson, 1999; McConnell and Perez-Quiros, 2000; Stock and Watson, 2002). Different from existing work, our study examines a much longer sample period and utilizes regionally disaggregated data.

Second, we extend existing work that constructs regional measures of economic activity for the United States and studies regional business cycles. Crone and Clayton-Matthews (2005), Aruoba, Diebold and Scotti (2009), Arias, Gascon and Rapach (2016), and Baumeister, Leiva-León and Sims (2024) construct economic activity indices for states (or MSAs), but their time series do not start until after the beginning of BEA's state-level GDP in 1963. Bokun et al. (2023) introduce a real-time database with 28 indicators per state for recent decades. We contribute to this literature by constructing new time series pre-dating the official statistics that have annual frequency, providing data on 60 indicators, and estimating an annual economic activity index that covers a much longer time span. Our analysis of state-level business cycles is related to existing work on state-level

business cycles including, among others, Owyang, Piger and Wall (2005), Owyang, Rapach and Wall (2009) and Hamilton and Owyang (2012). Our contribution is to extend such efforts by taking a historical perspective. In spirit, our work also builds on a growing strand of literature using regional identification for answering questions in macroeconomics (for a review of this literature, see Nakamura and Steinsson (2018)).

2 Data

In this section, we introduce our new state-level historical dataset that covers the 48 contiguous states (excluding Alaska, Hawaii, and Washington D.C.) for the period 1863-2021. Section 2.1 describes the data sources. Section 2.2 summarizes the variables included in our dataset. Section 2.3 provides details on how we construct the time series. Section 2.4 compares our dataset with existing work. A companion data appendix documents further details on the dataset.

2.1 Data Sources

Our data collection starts with two major publications compiled by the Census Bureau: The Statistical Abstract of the United States (henceforth referred to as SA) and the official decennial publications by the United States Census Bureau (henceforth referred to as Census). The SA is published on an annual basis starting from 1878, while the Census is published decennially starting from 1790. Drawn from various state and federal government reports, these two publications contain a wealth of state-level economic indicators.

However, much of the data contained in these publications has not been previously digitized, especially at the state level. This issue is especially pronounced for the SA, where state-level statistics are often not included in existing digitization efforts. We utilize Optical Character Recognition (OCR) technology as implemented by Amazon Textract to process the scanned documents, and then check for transcription errors with manual verification. Note that past data is frequently revised in later issues of the SA, based on revision by the agencies from whom the data is obtained. To account for this, we always use the data from the latest issue of the SA for which a given year's data is reported.

In some cases, data recorded in the SA or Census are presented in less detail than in the original underlying publications, or they do not span the entire length of our sample period. In an effort to construct a dataset that is as complete as possible, we draw upon a broader spectrum of historical

data sources, physical and digital, including government reports, books, private industry surveys, as well as previous work in the economic history literature. Much of this data is difficult to obtain and only available in print or PDF format. As a result, a major contribution of our work is to digitize many data sources previously not available in digital format.

The total number of sources we use is 113, of which 84 were newly digitized, while the remainder is compiled from scattered but already digitized sources. Section I in the supplementary data appendix provides a full list of all the variables together with their sources and coverage across states and time.[1]

Taken together, we provide a comprehensive and consistent set of state-level historical series that are comparable with their modern counterparts. They are not only the key inputs in the dynamic factor estimation we will focus on later, but will likely be of interest to researchers studying the economic history of U.S. states.

2.2 Main Variables

We focus on variables for which there are both modern-day equivalents and sufficient historical data. For example, since we are unable to identify a sufficient number of data points for retail sales (reported in SA) for the period before World War II, we do not include it in our dataset. That said, given our extensive research into historical publications and government reports containing state-level economic statistics, to the best of our knowledge, this is the most comprehensive state-level dataset that has ever been constructed for such a long time span. In fact, most variables have close to universal coverage, spanning from 1860s or 1870s until today. Some others, such as the number of motor vehicle registrations, are available starting from the early 1900s.

Our dataset contains a total of 60 individual variables, which can be grouped into seven broad categories: Real Activity, Government Finances, Labor Market, Transportation, Wealth, Housing, and Miscellaneous.

Real Activity. Our dataset covers real economic activity across three major sectors that are especially important in the earlier years of our sample: agriculture, mining, and manufacturing. The variables we construct include the value of agricultural products sold, the value of minerals, and the value added by the manufacturing sector. Within the agriculture and mining sectors, we

[1] For additional information on these data sources, we refer interested readers to the appendix, where we also include several examples of the tables in their original formats to highlight the challenge of extracting these data from many disparate sources that come in different formats.

collect data on major products, which are usually reported separately annually, and use them to estimate total values in these sectors whenever they are not reported on an annual basis in the early years.[2] We provide details on this process in Section 2.3.

In addition to sectoral output, we also report data on alternative cyclical indicators such as the number and liabilities of business failures, and the total number of business concerns, which have been recognized as important indicators of economic crises (Simpson and Anderson, 1957). The fact that they have been consistently reported since the late 19th century makes them especially suitable for long-run studies of the business cycle. Moreover, we report the value of imports and exports of merchandise, matched to states based on their customs district. Given that only some states have ports, we would expect these measures to matter for economic activity in certain states more than others.

Local consumption data has been notoriously difficult to obtain even for the post-war period. Nonetheless, our dataset attempts to construct some measure of expenditure in the historical context. In the US, expenditure on motor vehicles is known to be very sensitive to aggregate demand. For example, Orchard, Ramey and Wieland (2024) find that the marginal propensity to consume is 0.3 on motor vehicles and 0 on other consumption, suggesting that motor vehicle expenditure can be a key indicator for business cycles. While direct expenditure data are not available throughout our sample period, we include motor vehicle registrations, which are available since 1900, and automobile tax revenues, available from 1913.

Transportation. Given the importance of transportation networks in facilitating the flow of goods and people—and therefore, economic growth (e.g., Donaldson and Hornbeck, 2016)—our dataset includes measures of transportation, such as the mileage of railway tracks, rural roads, and state highways.

Government Finances. Our dataset reports several state-level fiscal variables on revenue, expenditure and debt. In particular, we include state government revenue, federal government internal revenue (as well as personal and corporate income tax revenues), state government total expenditure, and state government gross, net, and long-term debt. Wallis (2000) outlines the changing importance of the different levels of governments over time, and in particular the move from state and local funding to a federal system. Building upon Sylla, Legler and Wallis (1993), our dataset

[2]Examples of these major products include: the value of sheep, sweet potato crop and lumber produced, and the value of petroleum at mines, respectively.

includes regional variation in state and federal government activity with detailed personal and corporate income tax data.

Labor Market, Wealth, Bankruptcies, Miscellaneous. We cover measures of the labor market, including total non-farm employment, manufacturing employment, and manufacturing payroll, which allow us to track local economic dynamics via labor market fluctuations. We also report measures of personal income and the value of farmland and buildings. We extend the official BEA data on personal income that starts in 1929 back to 1880 at decennial intervals, and from 1919-1921 and in 1927-1928 annually. We also use several miscellaneous series. The number of bankruptcies includes both corporate and personal bankruptcies. Our banking sector data include bank assets, deposits, capital, liabilities and loans of national and state banks that stretch back to 1863, taken from Hoon et al. (2025). We report annual data on population starting from 1870, where we estimate the intercensal years by following the Census Bureau's technical reports. Finally, we report measures of patents, sentiments, newspaper circulation, as well as house and rental prices, the bulk of which draws upon existing work.

Our Data Appendix Table 1 tabulates a full list of variables, including their coverage across states and time, data sources, and frequency in the raw and imputed data.

2.3 Constructing Coherent Time Series

This section describes our approaches in constructing consistent and coherent state-level time series data.

Territorial Changes. Given the time span of our sample, many variables stretch back to before states were admitted to the Union in their current form. In order to ensure the data is comparable over time, we either combine or split state-level data. For example, data on the Oklahoma and the Indian Territory was reported separately in the raw data before they were jointly admitted to the Union in 1907. Accordingly, from 1870-1906, we report in our dataset the sum of both territories under "Oklahoma."

Consistency of Variable Definitions. Considering the length of the sample period and the breadth of the sources we draw on, we pay attention to maintaining consistent variable definitions across time and data sources. Whenever possible, we manually harmonize the raw data to account for definitional changes over time. This process typically entails checking source documents and

data files. For example, from 1921 onwards, the Annual Survey of Manufactures (ASM) stops collecting data on establishments with products valued between $500 to $5000. Since the Census of Manufacturing (CM) reports establishments by product value bin from 1905-1919, we are able to exclude establishments with products valued between $500 to $5000 before this change, such that the series remains comparable. Similarly, since the CM does not report data on the number of manufacturing establishments between 1947 to 1950, while the County Business Patterns (CBP) do, we impute the CM data using the CBP data using the same variable definitions.

As an illustration of this harmonization process, Panel (a) of Figure 1 displays our long-run series of harmonized value added of manufacturing production for New York and Texas. Another example in our dataset is the state government general revenue series shown in Panel (b) of Figure 1, where we combine multiple data sources and carefully verify variable definitions across time to ensure consistency.

When the nature of definitional changes is unclear, or when there is insufficient information to perform manual harmonization, we resort to ratio splicing the raw data from multiple sources. As an example, our coverage of the number of business failures series from Dun and Bradstreet ends in 1998. To extend the series to 2021, we ratio-splice the Dun and Bradstreet data with data on business bankruptcies collected from Hansen, Davis and Fasules (2016) for 1998-2007 and from US Bankruptcy Court reports for 2008-2021. We perform the ratio splice using overlapping data in 1998. Multiple other ratio splices are involved in constructing the business failure series, as illustrated in Panel (c) of Figure 1 for Ohio. Specifically, the imputed series incorporates four ratio splices for the pre-1934 data, three for 1934, two for 1935–38, one for 1939–1983, and two for 1984–96. Details of each ratio splice are provided in the companion data appendix.

Imputation. Our raw data series still remain incomplete after these changes, with most series containing missing data points that occur randomly, at regularly intervals, or both. For sporadic gaps involving only a single year, we use linear interpolation as a simple rule-of-thumb imputation method. For longer gaps—typically occurring at five- or ten-year intervals in the earlier years of our sample—we recover the missing observations through the factor model estimation, which produces these annual values as a by-product of the factor estimates.

An exception is the total output of the agricultural and mining sectors, where we estimate the low-frequency aggregate values using their annually-available underlying components. In particular, the value of agricultural products sold is only reported every ten years in the Census between

1870 to 1924, after which it is reported annually by the United States Department of Agriculture (USDA). Despite the absence of annual totals for much of this period, we have annual sales receipts data for major crop and livestock commodities covering 1870–1924. The aggregate of these individual receipts contains useful information regarding the fluctuations of the total value of agricultural products sold. We therefore use the growth rates of these individual receipts to impute the missing annual observations of the total value of agricultural products sold, following a constrained minimization approach in the spirit of Denton (1971). We describe the imputation details in Data Appendix 2.2, and display the results for California and Massachusetts in Panel (d) of Figure 1. We also conduct several robustness exercises, including one using the value-weighted growth of individual crops as proxies for growth in total agricultural products sold. We find that the imputed time series is fairly robust across these alternative approaches.

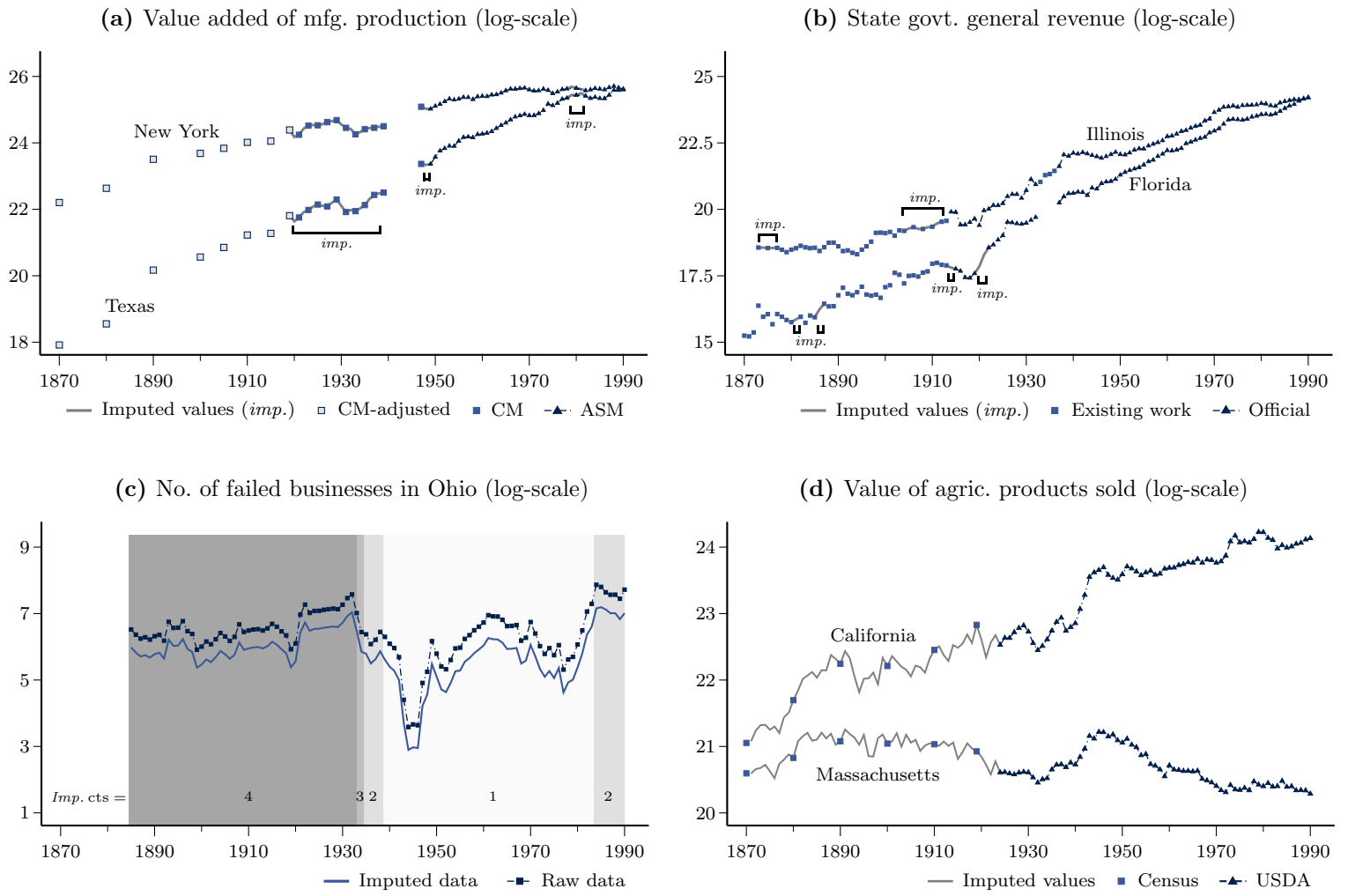
In total, we report 60 time-series variables that span economic activity across multiple sectors. To illustrate the structure and completeness of our dataset, Figure A.1 displays the fraction of variables available for each year by state, which underscores both the depth and breadth of our state-level panel that we construct.

2.4 Comparison with Existing Datasets

Table 1 compares our new dataset with existing state-level and U.S.-wide historical datasets capturing economic activity. For the former, our data provide an entirely new historical perspective, adding around 100 years of data that are useful in studying state-level economic dynamics from a long-run perspective. Additionally, the dataset we construct is much more comprehensive in terms of the number of indicators. In sum, we believe the dataset we construct is a significant addition to the existing literature in terms of length and breadth. To the best of our knowledge, there is no other data effort incorporating historical time series in a comparable manner.

Our data effort echoes works that attempt to build nationwide historical datasets. While we cannot directly compare our work to estimates of U.S. economic activity, it may be useful to compare their coverage. For example, Romer (1989) estimates Gross National Product (GNP) between 1869 and 1908. Davis (2004) estimates industrial production for the 1790-1915 period, just before the Federal Reserve's G.17 index of industrial production starts in 1919. Official U.S. GDP estimates from the BEA start in 1929. Different from these efforts, our dataset emphasizes a regional dimension and takes a "big data" approach by covering a large number of individual economic indicators.

Figure 1: Imputed and Harmonized Historical Time Series: Selected Examples



Notes: This figure displays examples of the constructed time series for selected states from 1870 to 1990. In Panel A, manufacturing value added is constructed using production output and cost data from the Census of Manufactures for 1870, 1880, 1890, 1900, 1905, 1910, 1915, 1947, and biennially for 1919–1939 (inclusive). Data pre-1921 are adjusted to match the product coverage of later years. Comparable output and cost data are drawn from the Annual Survey of Manufactures for 1949–1978, 1980, and the post-1982 years. Missing values for 1948, 1979, 1981, and biennial gaps for 1920–1938 are imputed using linear interpolation between adjacent years. In Panel B, the official sources for Illinois include the Financial Statistics of States (FSS; 1915–1919, 1921–1932, 1937–1950), State Government Finances (SGF; after 1950), and individual state reports (1914 and 1920). Non-official sources are from Hindman (2010) for 1873, 1875, 1877–1904, 1906, 1908, 1910, 1912, and 1913, and from Sylla, Legler and Wallis (1993) for 1933–1936. Missing values for 1874, 1876, and biennial gaps 1905–1911 are imputed using linear interpolation as before. For Florida, general revenue data are sourced from FSS and SGF for the same periods, with the exception for 1921, for which no records are available. Non-official sources are from Hindman (2010), covering 1870–1880, 1882–1885, and 1887–1913. Missing observations for 1881, 1886, and 1920–1921 are imputed via linear interpolation, while the 1914 observation is imputed based on implied growth rates from the 1913–1914 data in Sylla, Legler and Wallis (1993). In Panel C, the number of failed businesses is compiled from various Dun & Bradstreet (D&B) reports spanning the displayed period. In addition to the raw D&B data, we provide an imputed series with consistent variable definitions across the full sample. The imputation details are outlined in the Data Appendix, and the imputation counts (labeled *Imp. cts*) are reported in the plot for reference. Panel D shows the value of agricultural products sold, sourced from the USDA (yearly after 1923) and the Census (1870, 1880, 1890, 1900, 1910, and 1919), with the latter harmonized to match USDA definitions. Intercensal observations prior to 1924 are imputed using sales receipts from individual crop, livestock, and forest products. Finally, the values in Panels A, B, and D are in 2012 dollars, deflated using the U.S. price index from Williamson (2025).

Table 1: Comparison with Existing Datasets

	Variable	Frequency	Coverage
A. State-Level			
This paper	Economic activity index	Annually	1871–2021
BEA	Personal income	Annually	1929–2024
BEA	GDP	Annually	1963–2024
Crone and Clayton-Matthews (2005)	Coincident index	Monthly	1978–2003
Baumeister, Leiva-León and Sims (2024)	Economic conditions index	Weekly	1987–2023
B. National-Level Historical Data			
Davis (2004)	Industrial production	Annually	1790–1915
Miron and Romer (1990)	Industrial production	Monthly	1884–1940
Federal Reserve	Industrial production	Monthly	1919–2023
Williamson (2025)	GDP	Annually	1790–2023
Balke and Gordon (1989)	GNP	Annually	1869–1929
BEA	GDP	Annually	1929–2023

3 Estimating a State-Level Index of Economic Activity

In our dataset, variables are observed at varying frequencies—every ten, five, or two years, or every year. This mixture of frequencies occurs both across and within variables. For example, state-level manufacturing value added is available every ten years before 1910, every four to five years from 1910 through the 1920s, every two years until 1949, and annually afterwards. To take full advantage of the available data, we need an estimation framework that accommodates frequency variation across both the cross-section and time. We adopt the dynamic factor model framework of Baumeister, Leiva-León and Sims (2024) and modify it to accommodate the varying frequencies pertinent to our state-level dataset.[3]

3.1 Estimation Framework

Following Baumeister, Leiva-León and Sims (2024), we postulate that there is a latent stationary factor, $f_{i,t}$, that is common to N_i observable indicators for state i. We model the common factor as an annual series, with $t = 1, 2, \ldots, T$ indexing individual years over our sample period. For each state i, let N_i represent the total number of indicators used in the estimation. Of these indicators,

[3] In Baumeister, Leiva-León and Sims (2024), the authors construct state-level economic conditions indices based on indicators with weekly, monthly, and quarterly reporting frequencies. Similar to Baumeister, Leiva-León and Sims (2024), earlier studies by Crone and Clayton-Matthews (2005), Aruoba, Diebold and Scotti (2009), and more recently Lewis et al. (2022) consider time series sampled at different frequencies to construct economic coincidence indices within a dynamic factor model framework. Earlier studies, including works by Stock and Watson (1989, 1991), consider indicators with one frequency.

let N_i^y denote those that report only at annual frequency, and let $N_i - N_i^y$ denote those that report at mixed frequencies. We let the corresponding sets of indicators be represented by $\gamma(N_i)$, $\gamma(N_i^y)$, and $\gamma(N_i - N_i^y)$, respectively. For each indicator $j \in \gamma(N_i)$, let $Y_{i,j,t}$ denote its value for year t. If $j \in \gamma(N_i^y)$ and $Y_{i,j,t}$ is reported in levels, we compute j's annual growth rates using log-differences such that $y_{i,j,t} = \ln Y_{i,j,t} - \ln Y_{i,j,t-1}$; if $Y_{i,j,t}$ is reported in growth rates, we simply set $y_{i,j,t} = Y_{i,j,t}$. We assume that $y_{i,j,t}$ is associated with $f_{i,t}$ through the following structure:

$$y_{i,j,t} = \lambda_{i,j} f_{i,t} + u_{i,j,t}, \tag{1}$$

where $\lambda_{i,j}$ denotes the factor loading of indicator j to $f_{i,t}$. $u_{i,j,t}$ is an idiosyncratic factor, capturing idiosyncratic variations of indicator j. We assume $f_{i,t}$ follows a Gaussian AR($l_{i,f}$) process and $u_{i,j,t}$ follows a Gaussian AR($l_{i,u}$) process given by:

$$f_{i,t} = \phi_{i,1} f_{i,t-1} + \phi_{i,2} f_{i,t-2} + \cdots + \phi_{i,l_{i,f}} f_{i,t-l_{i,f}} + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim N(0, \sigma_{i,f}^2), \tag{2}$$

$$u_{i,j,t} = \psi_{i,j,1} u_{i,j,t-1} + \psi_{i,j,2} u_{i,j,t-2} + \cdots + \psi_{i,j,l_{i,u}} u_{i,j,t-l_{i,u}} + \varepsilon_{i,j,t}, \qquad \varepsilon_{i,j,t} \sim N(0, \sigma_{i,j}^2). \tag{3}$$

Following standard practice in dynamic factor model estimation, we fix the scale of the autoregressive coefficients in equation (2) by setting $\sigma_{i,f} = 1$ for all i. Moreover, we normalize $y_{i,j,t}$ to have zero mean and unit variance before estimation. The former removes the need for a constant term in equation (1). While unit-variance is not necessary for identification, it can be convenient for interpretation; see Crone and Clayton-Matthews (2005, p. 594) for a discussion.

Suppose indicator $j \in \gamma(N_i - N_i^y)$. Then, the indicator has mixed reporting frequencies over the sample period. Let $\mathcal{T}_{i,j,t} \geq 1$ denote the number of years since indicator j was last reported in year t. For instance, if indicator j is observed in 1880 and 1890 for state i, then $\mathcal{T}_{i,j,1890} = 10$. Note that $\mathcal{T}_{i,j,t}$ varies over time for $j \in \gamma(N_i - N_i^y)$ to account for its mixed reporting frequencies. Now, let $z_{i,j,t}$ be an auxiliary variable denoting the annual growth rates of indicator j. If $\mathcal{T}_{i,j,t} = 1$, then $z_{i,j,t} = y_{i,j,t}$; otherwise, $z_{i,j,t}$ is unobserved if $\mathcal{T}_{i,j,t} > 1$. Using $z_{i,j,t}$ allows us to express indicator j's annualized growth rates between years t and $t - \mathcal{T}_{i,j,t}$ in terms of $f_{i,t}$ as follows:

$$\begin{aligned}
&\frac{1}{\mathcal{T}_{i,j,t}} \left(\ln Y_{i,j,t} - \ln Y_{i,j,(t-\mathcal{T}_{i,j,t})} \right) \\
&= \frac{1}{\mathcal{T}_{i,j,t}} \left(z_{i,j,t} + z_{i,j,t-1} + \cdots + z_{i,j,(t-\mathcal{T}_{i,j,t}+1)} \right) \\
&= \frac{1}{\mathcal{T}_{i,j,t}} \lambda_{i,j} \left(f_{i,t} + f_{i,t-1} + \cdots + f_{i,(t-\mathcal{T}_{i,j,t}+1)} \right) + \frac{1}{\mathcal{T}_{i,j,t}} \left(u_{i,j,t} + u_{i,j,t-1} + \cdots + u_{i,j,(t-\mathcal{T}_{i,j,t}+1)} \right),
\end{aligned} \tag{4}$$

where the final equality follows from equation (1). The above derivation effectively expresses the (annualized) growth rates of all indicators in $\gamma(N_i - N_i^y)$ as lag polynomials of the common factor and idiosyncratic term. Equations (1) and (4) together constitute the observation equation in the state-space representation of the following section.

State-Space Representation. We can express equations (1) and (4), along with equations (2) and (3), in a Gaussian state-space structure:

$$\mathbf{y}_{i,t} = \mathbf{H}_{i,t}\boldsymbol{\alpha}_{i,t}, \tag{5}$$

$$\boldsymbol{\alpha}_{i,t} = \mathbf{T}_i\boldsymbol{\alpha}_{i,t-1} + \boldsymbol{\eta}_{i,t}, \qquad \boldsymbol{\eta}_{i,t} \sim N(\mathbf{0}, \mathbf{Q}_i), \tag{6}$$

for $t = 1, \ldots, T$. In equation (5), $\mathbf{y}_{i,t}$ is a column vector of length $n_{i,t}$ that collects the observed growth rates in year t for state i. We note that $n_{i,t} \leq N_i$, and the inequality is strict when there are missing values in year t. $\boldsymbol{\alpha}_{i,t}$ is the state vector, given by:

$$\boldsymbol{\alpha}_{i,t} = \Bigg[\Upsilon_{c_{i,1}}(L)f_{i,t},\ \underbrace{\Upsilon_{c_{i,1}}(L)u_{i,1,t},\ \ldots,\ \Upsilon_{c_{i,N_i-N_i^y}}(L)u_{i,N_i-N_i^y,t}}_{N_i - N_i^y \text{ terms with indicator } j \in \gamma(N_i - N_i^y)},\ \underbrace{\Upsilon_{l_{i,u}}(L)u_{N_i-N_i^y+1,t},\ \ldots,\ \Upsilon_{l_{i,u}}(L)u_{N_i,t}}_{N_i^y \text{ terms with indicator } j \in \gamma(N_i^y)}\Bigg]^\top,$$

where $\Upsilon_{c_{i,j}}(L)$ defines a vector of lag operators given by:

$$\Upsilon_{c_{i,j}}(L) = \left(L^0,\ L^1,\ L^2,\ \ldots,\ L^{\max_t(\mathcal{T}_{i,j,t})-1}\right), \quad \text{for all } t = 1, \ldots, T.$$

We order the indicators in $\mathbf{y}_{i,t}$ such that:

$$\Upsilon_{c_{i,1}}(L) = \left(L^0,\ L^1,\ L^2,\ \ldots,\ L^{\max_{j,t}(\mathcal{T}_{i,j,t})-1}\right), \quad \text{for all } t = 1, \ldots, T \text{ and } j \in \gamma(N_i).$$

Likewise, we define:

$$\Upsilon_{l_{i,u}}(L) = \left(L^0,\ L^1,\ L^2, \ldots,\ L^{l_{i,u}-1}\right),$$

where $l_{i,u}$ denotes the number of autoregressive lags in equation (3). L denotes the lag operator, such that $L^k x_t = x_{t-k}$ for a variable x_t. We note in passing that the length of $\boldsymbol{\alpha}_{i,t}$ can be computed as $\max_{j,t}(\mathcal{T}_{i,j,t}) + \sum_{j=1}^{N_i - N_i^y} \max_i(\mathcal{T}_{i,j,t}) + N_i^y \times l_{i,u}$, for $t = 1, \ldots, T$ and $j \in \gamma(N_i)$. Now, matrix $\mathbf{H}_{i,t}$ has $n_{i,t}$ rows by construction. The j-th row of $\mathbf{H}_{i,t}$ consist of $\lambda_{i,j}$, $\mathcal{T}_{i,j,t} \geq 1$, and possibly zeros,

so that equation (4) holds in the j-th row of equation (5). Likewise, matrix $\mathbf{T}_i$ and vector $\boldsymbol{\eta}_{i,t}$ are parameterized such that equation (6) stacks the autoregressive processes in (1) over all indicators in $\gamma(N_i)$. Unlike $\mathbf{H}_{i,t}$, $\mathbf{T}_i$ is not time-varying since the autoregressive orders $l_{i,f}$ and $l_{i,u}$ are fixed in our estimation.[4]

Estimation Strategy. For notational simplicity, we omit the index i from equations (5)–(6). The state-space system is estimated using a Bayesian MCMC approach via Gibbs sampler. The procedure is standard: in each MCMC iteration, we first obtain a draw of the state vector conditional on the model parameters and the full information set $\mathcal{F}_T \equiv (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_T^\top)^\top$. Then, conditioning on the draw of the state vector and the observations $\mathcal{F}_T$, we update the model parameters. Further details on the estimation algorithm and the assumed priors are described in Appendix B.

Backing out the Economic Activity Index. For state i, we follow Baumeister, Leiva-León and Sims (2024) to approximate the index of economic activity using the following equation:

$$\tilde{f}_i = (\boldsymbol{\lambda}_i^\top \boldsymbol{\lambda}_i)^{-1} \boldsymbol{\lambda}_i^\top \mathbf{y}_i^P, \tag{7}$$

with $\tilde{f}_i \equiv [\tilde{f}_{i,1}, \tilde{f}_{i,2}, \ldots, \tilde{f}_{i,T}]^\top$. $\boldsymbol{\lambda}_i$ is an $(N_i \times 1)$ vector containing the median estimates of the factor loadings. Moreover, $\mathbf{y}_i^P \equiv [\mathbf{y}_{i,1}^P, \mathbf{y}_{i,2}^P, \ldots, \mathbf{y}_{i,T}^P]$ is the $(N_i \times T)$ input data with missing observations replaced by the projected values of the Kalman filter. According to Baumeister, Leiva-León and Sims (2024, p. 488), using $\tilde{f}_i$ in place of $\tilde{f}_i$ provides two advantages: (i) while the two measures are typically close across time, the former minimizes the effect of revisions to the factor estimates when new information is added, and (ii) the contribution of the jth input series to $\tilde{f}_{i,t}$ can be conveniently computed as $(\boldsymbol{\lambda}_i^\top \boldsymbol{\lambda}_i)^{-1} \lambda_{i,j} y_{i,j,t}^P$. Because of the identification assumptions in the estimation, as well as the normalization of the input indicators in $\mathbf{y}_i^P$, $\tilde{f}_i$ needs to be rescaled in some ways to ensure that it is interpretable as an index for the state's economic activity. We follow Clayton-Matthews and Stock (1998) and scale $\tilde{f}_i$ so that the resulting index, from time period 1964 to 2021, has an average growth and variance matching those of the state's real GDP growth rates during the same period. More specifically, the scaled index of economic activity for state i is obtained by the

[4] In our baseline estimation, we choose $l_{i,f} = l_{i,u} = 4$ for all i to align the autoregressive lag length with the average peak-to-peak business cycle duration (3.9 years), as measured by the NBER since the early 1880s. We consider $l_{i,f} = l_{i,u} = 5$ in our sensitivity analysis so as to match the median peak-to-peak business cycle duration (4.9 years).

following affine transformation:

$$s_{i,t} = \beta_{1,i} + \beta_{2,i}\tilde{f}_{i,t}, \qquad \text{for} \quad t = 1, 2, \ldots, T, \tag{8}$$

with $\beta_{1,i} = -\frac{\sigma_i}{\sigma_{\tilde{f}_i}} \times \mu_{\tilde{f}_i} + \mu_i$ and $\beta_{2,i} = \frac{\sigma_i}{\sigma_{\tilde{f}_i}}$. μ_i represents the average growth rates of state i's real GDP (in 2012 dollars) from 1964 to 2021. $\mu_{\tilde{f}_i}$ denotes the average value of $\tilde{f}_{i,t}$ from 1964 to 2021. σ_i and $\sigma_{\tilde{f}_i}$ are the standard deviations of state i's real GDP growth rates and $\tilde{f}_{i,t}$ over 1964–2021.

3.2 Estimation Results

Estimation Inputs. Our model is estimated for each of the 48 contiguous U.S. states separately. We select state-level indicators for our baseline estimation based on two criteria. First, we choose indicators that are economically relevant and tend to comove with business cycles. Second, we include indicators with long historical coverage and few missing observations in the early years of the sample. While our method is flexible and can accommodate missing observations, choosing indicators with more complete coverage helps improve the estimation's accuracy. In Table 2, we present the variables that together form the baseline inputs, together with their available period, frequency and geographic coverage. Our selected dataset covers series of varying frequencies—annual, 5-yearly and 10-yearly, sometimes varying within each variable. As detailed in Section 3.1, the flexibility of our model allows us to accommodate these variations in frequency.

Figure 2 plots our factor estimates against BEA GDP growth rates (available post 1963) for a few selected states. We observe a strong linear relationship between the factor estimates and GDP growth, suggesting that our estimated factor effectively captures economic activity at the state level. This observation validates the linear transformation in equation (8) in generating an index that is comparable to the familiar GDP growth.

The State Economic Activity Index. Using the baseline input variables in Table 2, we estimate a state-level economic activities index (SEAI) for each state at the annual frequency. Figure 4 shows our results in a heat plot that reveals distinct patterns of economic growth and contraction across different time periods and regions.

Several major downturns stand out, particularly the Great Depression of the 1930s, which was the most severe and widespread economic collapse in US history. States reliant on manufacturing, such as Michigan, Pennsylvania, and Ohio, experienced deep recessions, while agricultural states like

Table 2: Input Series Used in the Baseline Estimation

	Indicators	Geographic and temporal coverage	
		No. of states	Years covered
Real activity	Nonfarm employment	48	1880, 1890, 1900, 1910, 1920, 1929–2021
	Liabilities of failed firms	48	1886–1983
	Value of mining production	48	1881–2021
	Value of agri. products sold	48	1871–2021
	Value of exported merchandise	27	1872–1948, 1951-1952, 1955–1981, 1984–2021
	Value of imported merchandise	33	1872–1948, 1951-1952, 1955–1981, 1984–2021
	Value added of mfg. production	48	1880, 1890, 1900, 1905, 1910, 1915, 1920–2021
Wealth	Personal income	48	1890, 1900, 1910, 1920, 1928–2021
	Value of farmland and buildings	48	1910–2021
Govt.	State govt. gross debt	48	1871–2021
	State govt. general revenue	46	1871–2021
	Federal govt. internal revenue	48	1871–2021
Others	Housing sales price index	21	1891–2021
	Housing rental price index	21	1891–2006
	Railroad operating mileage	48	1871–1973
	No. of motor vehicle registration	48	1901–2021

Notes: This table lists the inputs included in the baseline estimation, with all inputs expressed as annual or annualized growth rates calculated using log-differences. Liabilities of failed firms and values of imports and exports are smoothed with a three-year moving average. The final column presents the years in which the inputs are available for at least one state. Intercensal values of agricultural products sold from 1871 to 1924 are imputed using annual growth rates from 16 major crop and livestock commodities; see Section 2.2 of the companion data appendix for imputation details.

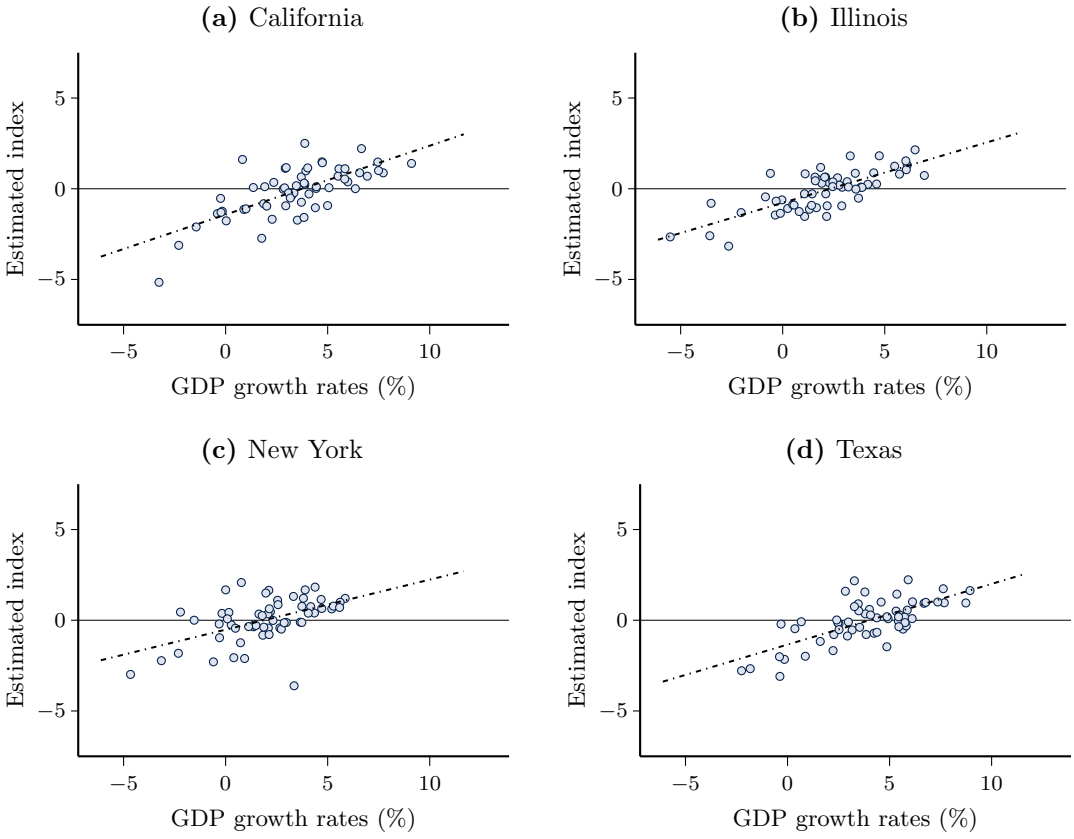
Oklahoma, Kansas, and Nebraska suffered due to the Dust Bowl. Thanks to the wide coverage of historical data, we also capture earlier recessions, including the Long Depression (1873-1896) and the Panic of 1893, that show significant declines particularly in railroad-dependent and farming states. More recent recessions, such as the 2008 Great Recession and the COVID-19 downturn of 2020, also display nationwide impacts, with financial hubs (New York) and real estate-heavy states (Florida, Arizona, Nevada) suffering severe contractions.

Periods of strong economic growth are equally evident. The post-World War II boom from the 1940s to the 1960s saw widespread economic expansion across most states, likely driven by industrial production, infrastructure development, and demographic growth. The 1990s also mark a period of significant economic expansion, largely due to the rise of the technology sector, benefiting states like California, Washington, and Massachusetts.

Our heat map also delivers a clear message regarding cross-state variation, with some states experiencing frequent boom-bust cycles while others show long-term stability.[5] Energy-dependent

[5]Table C.1 in the appendix offers a complementary perspective to the heat map by presenting descriptive statistics on

Figure 2: Factor Estimates v.s. GDP Growth Rates



Notes: This figure displays the association between the factor estimates (in standard deviations from zero) and annual GDP growth rates (in percentages) for selected states from 1964 to 2021. GDP data are from the BEA.

states such as North Dakota, Wyoming, and West Virginia exhibit high volatility, likely due to the highly volatile commodity prices important to their local economies. Similarly, states with large tourism and real estate sectors, such as Nevada and Florida, show sharp declines during financial crises but rapid recoveries during periods of expansion. In contrast, states like California, Texas, and New York demonstrate relatively consistent growth due to their diversified economies. The Rust Belt states, including Ohio, Michigan, and Pennsylvania, show prolonged periods of economic decline in the late 20th century due to the decline of manufacturing industries.

Over time, the structure of economic cycles has changed. Before 1950, recessions were longer and recoveries slower, often concentrated in specific regions. After World War II, economic downturns became shorter and recoveries faster, potentially mitigated by monetary policy, government stimulus, and broader economic diversification. Overall, this heat map illustrates the evolving na-

the estimated economic activity indices across states, providing static evidence on cross-state variation.

ture of the U.S. economy, highlighting how national economic cycles, industrial shifts, and policy changes shape state-level growth patterns.

Comparison with Existing State-Level Data. In order to validate that our estimates capture state-level business cycles, we compare them with existing measures that are available for a shorter period of time in a binscatter plot Figure 3. These data include: (i) state GDP from the BEA; (ii) the State Coincident Index from Philadelphia Fed;[6] (iii) the state-level unemployment rate from the BLS Local Area Unemployment Statistics; and (iv) state-level personal income from the BEA. As shown before, our factor estimates line up well with GDP, so it is not surprising that a linearly-transformed version also strongly correlates with GDP, displayed in Panel (a) of Figure 3. Similarly, the SEAI exhibits a strong correlation with established economic indicators, including personal income, State Coincident Indexes, and the unemployment rate. This consistency supports the validity of our index in capturing economic fluctuations over an extended period.

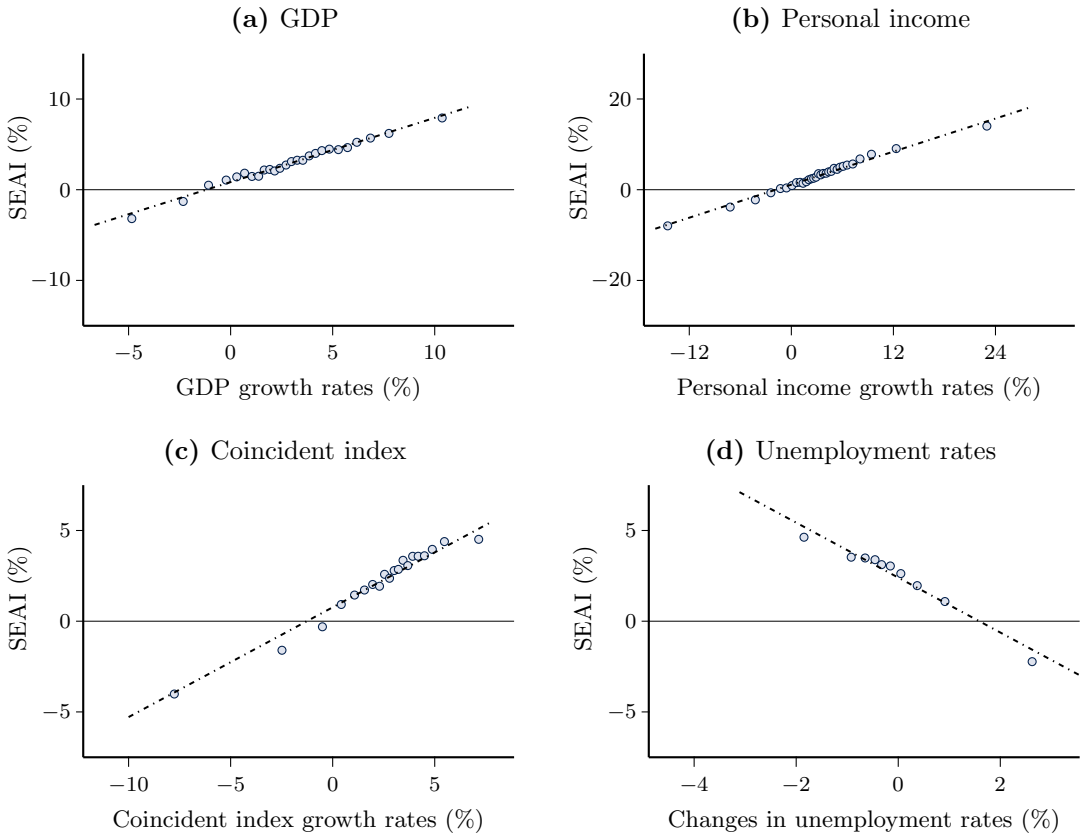
Alternative Input Specifications. We examine two alternative input specifications. The first specification augments the baseline set of indicators listed in Table 2 with three additional series: (i) total bank assets and liabilities from Hoon et al. (2025), and (ii) changes in lagged sentiment indicators from Van Binsbergen et al. (2024). The first two series capture dynamics in the financial sector, while the sentiment series provides a proxy for local economic expectations. As shown in Figure C.2, including these additional indicators does not materially alter the qualitative features of the estimated indices. State-level indices obtained under this extended specification remain highly correlated with their baseline counterparts, particularly in the case of New York.

As a further robustness check, we implement a data-driven selection of input indicators using ridge regression. For each state, we regress personal income—an independent measure of economic activity—on a large pool of potential input variables, including those in the extended baseline and additional candidates listed in Table 4. We explore 1,000 logarithmically spaced values for the ridge penalty parameter between 10^{-4} and 10^{4}, compute the corresponding slope coefficients, and use their absolute averages to evaluate indicator relevance. Specifically, we retain indicators whose average absolute coefficients exceed the 30th percentile across all regressors. We then use this subset of input indicators to re-estimate the state-level economic activity indices. Despite relying on a smaller and automatically selected set of inputs, the ridge-based estimates remain highly consistent

[6] https://www.philadelphiafed.org/surveys-and-data/regional-economic-analysis/state-coincident-indexes

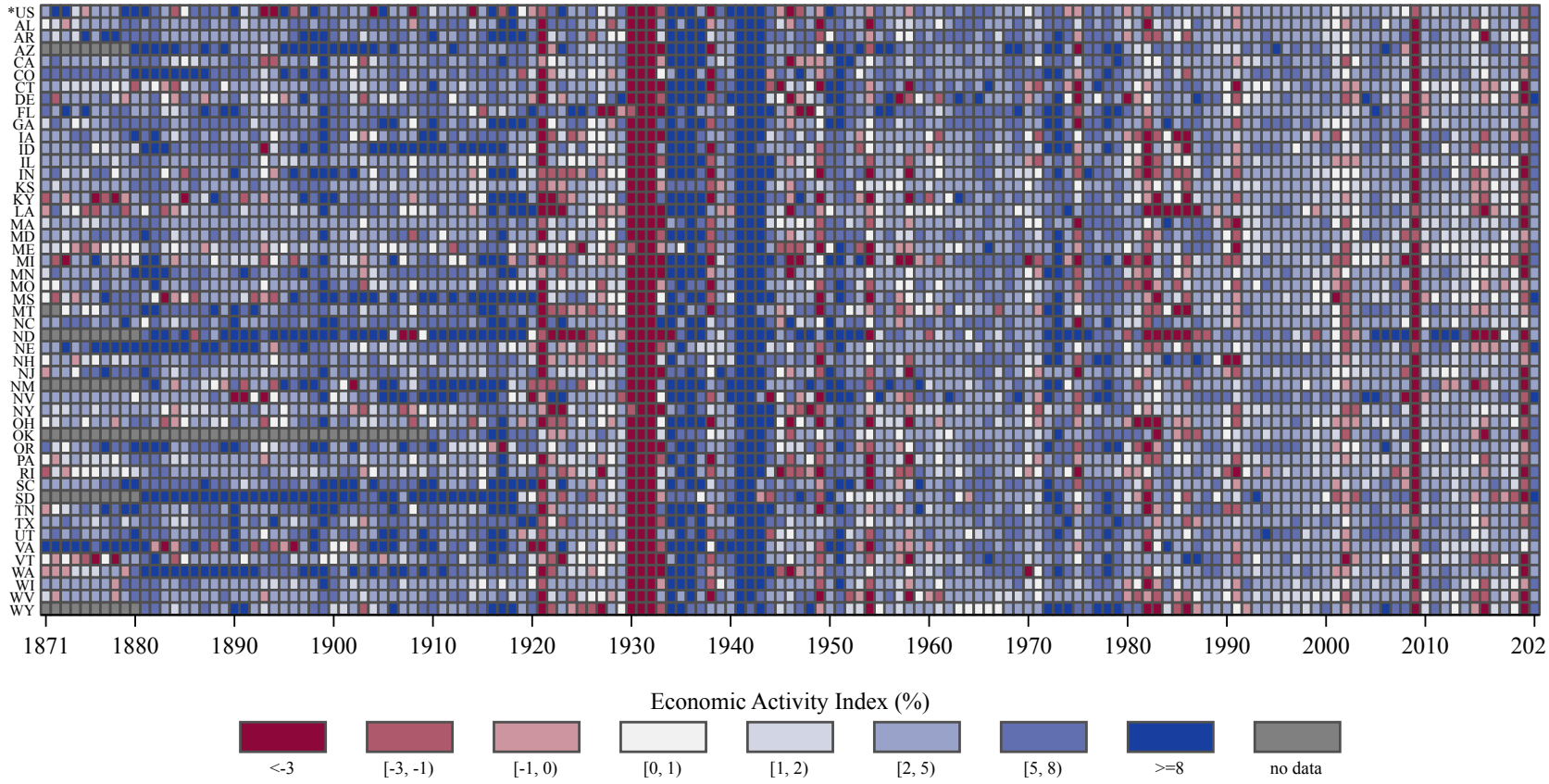
with those from the baseline specification, reinforcing the robustness of our approach to alternative input choices.

Figure 3: SEAI and Other Measures of Economic Conditions



Notes: This figure presents binscatter plots of the estimated economic activity indices against alternative measures of state-level economic conditions. The number of bins is chosen using the rule-of-thumb bin selector of Cattaneo et al. (2024). Annual growth rates of state-level GDP (1964–2021), personal income (1929–2021), and the coincident index (1980–2021) are calculated as log differences, while changes in unemployment rates (1977–2021) are computed as first differences. GDP and personal income data are from the BEA, coincident indices are from the Philadelphia Fed, and unemployment rates are from the BLS.

Figure 4: State-Level Indices of Economic Activity



Notes: Each cell represents the estimated state-level index of economic activity (in percentages). Gray cells indicate years for which the index is not estimated, often due to limited data availability before statehood. For reference, the first row reports US GDP growth rates from Williamson (2025), labeled as "*US".

4 150 Years of State-Level Business Cycles

Our estimated index provides various novel insights into state-level business cycles from a very long-run perspective. In this section, we discuss the drivers of our estimated index, the time-varying nature of state business cycles and their relationship with the national one, and the broader fluctuations in alternative economic indicators over the state business cycles.

4.1 Decomposition of the Estimated State Economic Activity

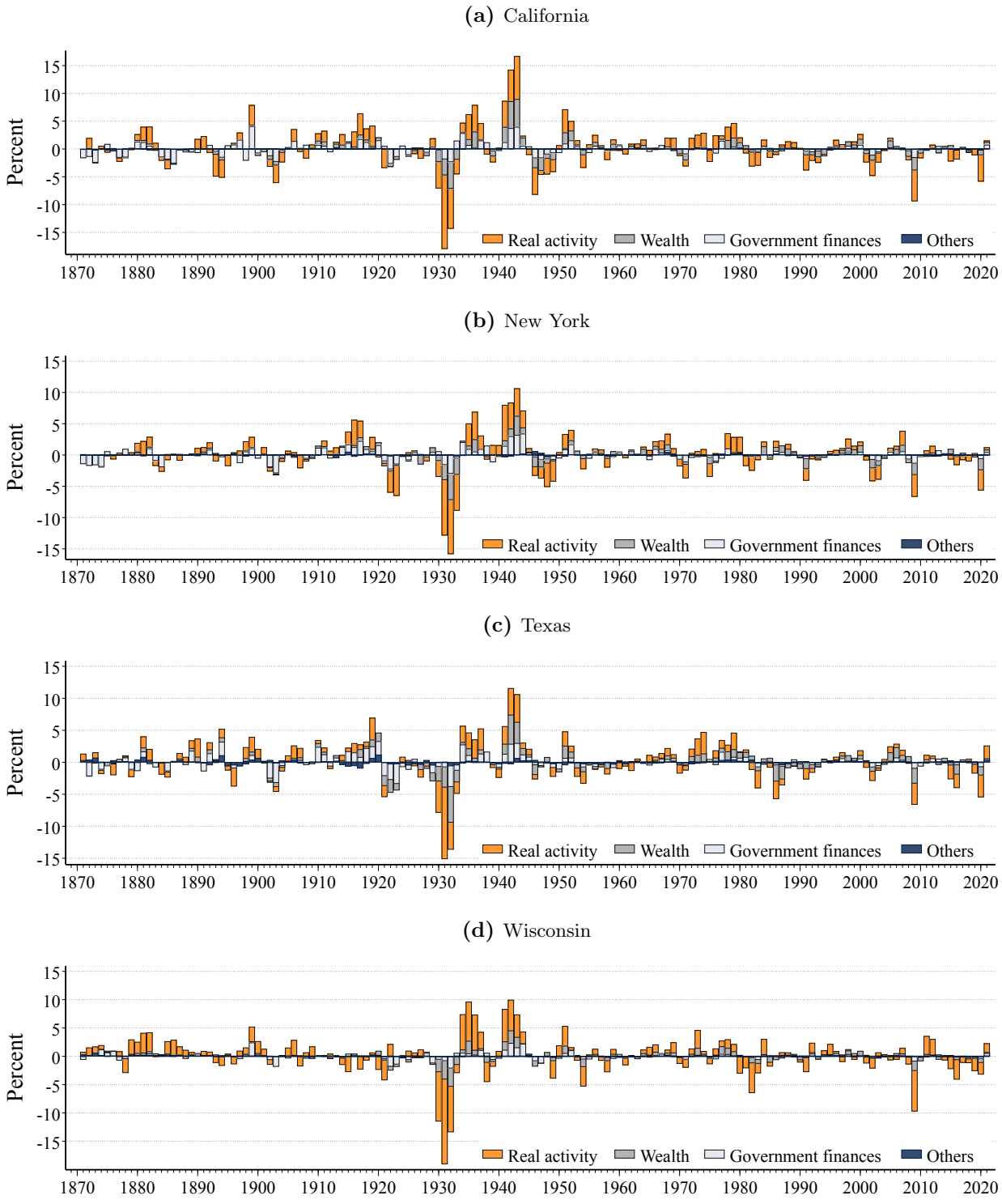
Our factor model allows for a decomposition of the estimated index into the factors that contribute to its variation over time. Figure 5 plots changes in economic activity for a selection of states together with changes in the underlying indicators grouped into four buckets: real activity, wealth, government finances, and others. This decomposition reveals several interesting patterns. In general, and perhaps not surprisingly, variables referring to real activity are the primary contributors to the estimated index both across states and over time. Other variables in the categories of government finances and wealth play a relatively more important role only in the earlier sample, but with considerable heterogeneity across states. For example, government finances play an important role in shaping California and New York's pre-1920 economic dynamics compared to the other three states, while in Wisconsin, fluctuations in variables directly related to real activity appear to be the primary drivers.

4.2 Heterogeneity, Volatility and Synchronization

In Figure 6, we present graphs of the estimated annual economic activity indices for selected states from 1871 to 2021, overlaid by NBER recession bars shaded in gray. This figure reveals similarities, but also key differences in the state-level business cycles, both within and across different regions. For example, for the states of Florida and Texas in the South, the economic indexes show strong comovement over the entire sample period, but they also exhibit very different dynamics in specific scenarios such as the post-World War II recovery period and the Great Recession. With the exception of the Great Depression, New Deal, and World War II, state-level business cycles do not appear to be more volatile before and after 1960, an observation consistent with that of Romer (1986).

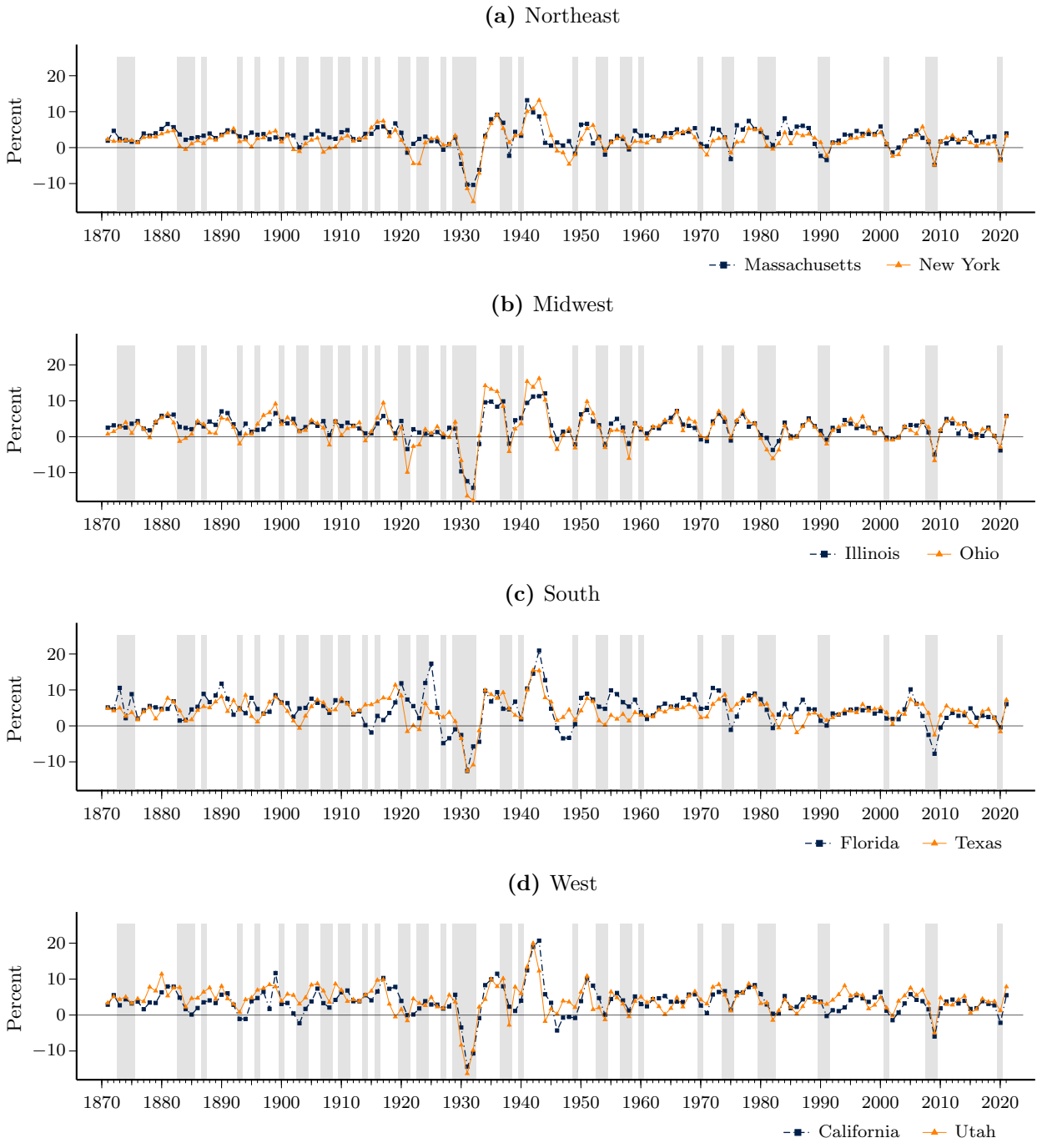
However, we observe a dramatic variation over time in the synchronization of the estimated economic activity indices across states. Figure 7 shows this by plotting the cross-state standard

Figure 5: Decomposition of Economic Activity Indices for Selected States



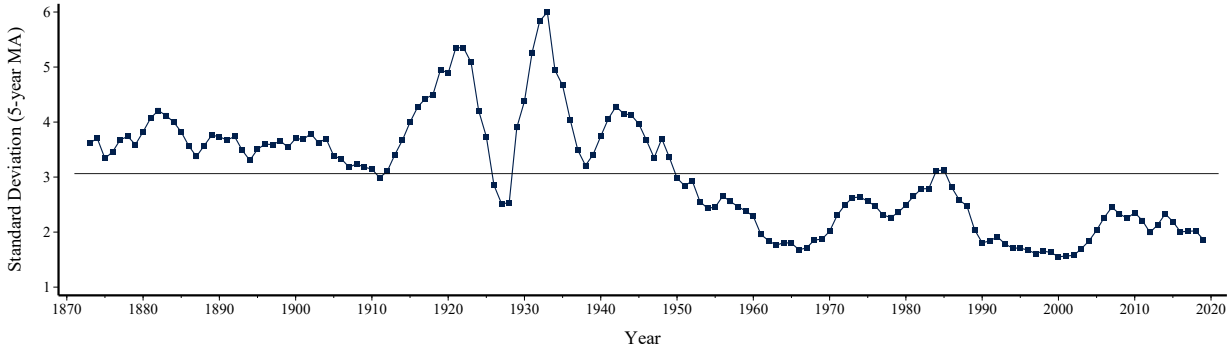
Notes: This figure decomposes the variation in the economic activity index for selected states into four categories: real activity, wealth, government finances, and others. The specific input series within each category are detailed in Table 2. Note that the economic activity index is normalized to have a mean of zero in this figure.

Figure 6: Annual Index of Economic Activity for Selected States



Notes: This figure displays the annual economic activity indices for selected states from 1871 to 2021. The shaded bars indicate recession years. Recession years from 1887 to 1991 are defined based on Table 3 of Romer (1999), with a year counted as a recession year if it reports at least one quarter within the peak-to-trough phase. Recession years prior to 1887 are defined according to Table 1 of Davis (2006), with a year counted as a recession year if it falls within the peak-to-trough phase. For years after 1991, the NBER chronology is used.

Figure 7: Dispersion of State-Level Economic Activities Over Time



Notes: This figure shows the standard deviation of our estimated economic activity indices across states as a proxy for business cycle synchronization, averaged over a five-year moving window. The horizontal line represents the average value over time, which is about 3.

Table 3: Dispersion of State-Level Economic Activity Index before and after WWII

	Pre-WWII		Post-WWII	
	1871–1905	1906–1940	1945–1980	1981–2019
All years	3.72	4.21	2.61	2.19
Recession years	3.69	4.40	2.71	2.61
Recession years, except the Great Depression	3.69	3.99	2.71	2.61
Non-recession years	3.74	4.01	2.57	2.09

Notes: This table shows the average dispersion of economic activity indices across states before and after WWII. Annual dispersion of economic activity is measured by standard deviations. For the definition of recession years, refer to the notes in Figure 6.

deviation of changes in economic activity over time. We find that the cross-state standard deviation in economic activity is notably higher in the pre-World War II period compared to afterwards, and that the Great Depression and to a lesser extent the 2007-08 Global Financial Crisis saw a dramatic increase in this measure. Table 3 further illustrates this pattern by summarizing the average dispersion of the state-level economic activity indices across four periods: 1871–1905, 1906–1940, 1945–1980, and 1981–2019. This exercise also shows a steady increase in business cycle synchronization after World War II, and this increase does not seem to be specific to expansion and recession periods. In Appendix C.1, we develop an alternative measure of business cycle synchronization based on Kalemli-Özcan, Papaioannou and Peydró (2013), which yields similar conclusions.

To sum up, our estimated historical state-level index of economic activity shows three interesting facts: (1) state cycles exhibit substantial heterogeneity; (2) state-level business cycle volatility has not changed much over the past 150 years, excluding the Great Depression; and (3) state-level

cycles have become more synchronized after World War II.

4.3 The Correlates of State Business Cycles

Business cycle research typically explores the extent to which fluctuations in GDP co-move with changes in a range of economic indicators. We compare our economic activity index against select variables *not* included in the index's inputs by running simple bivariate panel regressions with state fixed effects. In particular, we regress changes in state-level economic activity on the log-difference of several indicators, and then report the resulting coefficients, t-statistics, and (within-) R^2. We cluster standard errors by state.

Table 4 plots the results. The economic activity index we construct is highly correlated with changes in manufacturing payroll and employment, with an R^2 upwards of 0.2 and t-statistics exceeding 10. Business failures and bankruptcies are also highly predictive of changes in economic activity. We also find that a measure of state-level sentiment from Van Binsbergen et al. (2024) predicts economic activity over the more than 150-year period we consider, consistent with Van Binsbergen et al. (2024), who show a similar result for state-level GDP growth after 1963.

Importantly, we also find some indirect evidence suggesting that our measure of economic activity is correlated with output in the tertiary sector. Changes in the circulation of newspapers, which captures variation in part of the services sector that is historically important, attract a t-statistic of 4.70, suggesting a strong statistical link with our index. We also draw on the long-run historical data on patenting activity from Berkes (2018), and find a positive correlation with economic activity, suggesting strong procyclicality of innovation activities.

4.4 State and National Business Cycles

Does a national recession necessarily mean a recession happens in all states at the same time? Are some states experiencing upswings or downturns in the absence of major U.S.-wide business cycle events? We provide some new systematic evidence on these questions based on our large historical sample.

As discussed earlier, state-level business cycles are far from perfectly coinciding, although they seem to become more so during national downturns. The recessions in 1873, 1929, and 2007 in particular stand out for how widespread the regional economic downturns were. In contrast, the 1991 and 2001 recessions were much more concentrated in certain states. Figure C.5 plots the change in our economic activity index across states during three major nationwide recessions:

Table 4: Association of State-Level Economic Activity and Other Economic Indicators

Indicator	$\hat{\beta}$	t-stat	Within-R^2
Manufacturing payroll	8.31^{***}	11.14	0.20
Number of manufacturing employees	9.97^{***}	11.64	0.28
Number of manufacturing establishments	4.67^{***}	9.90	0.02
Number of patents	1.07^{***}	3.80	0.00
Number of bankruptcies commenced	-3.33^{***}	-7.85	0.03
Number of bankruptcies terminated	-1.28^{***}	-5.25	0.00
Number of business failures	-4.98^{***}	-10.61	0.06
Total circulation of newspapers	3.55^{***}	4.70	0.03
Change in lagged sentiments	1.81^{***}	6.79	0.01

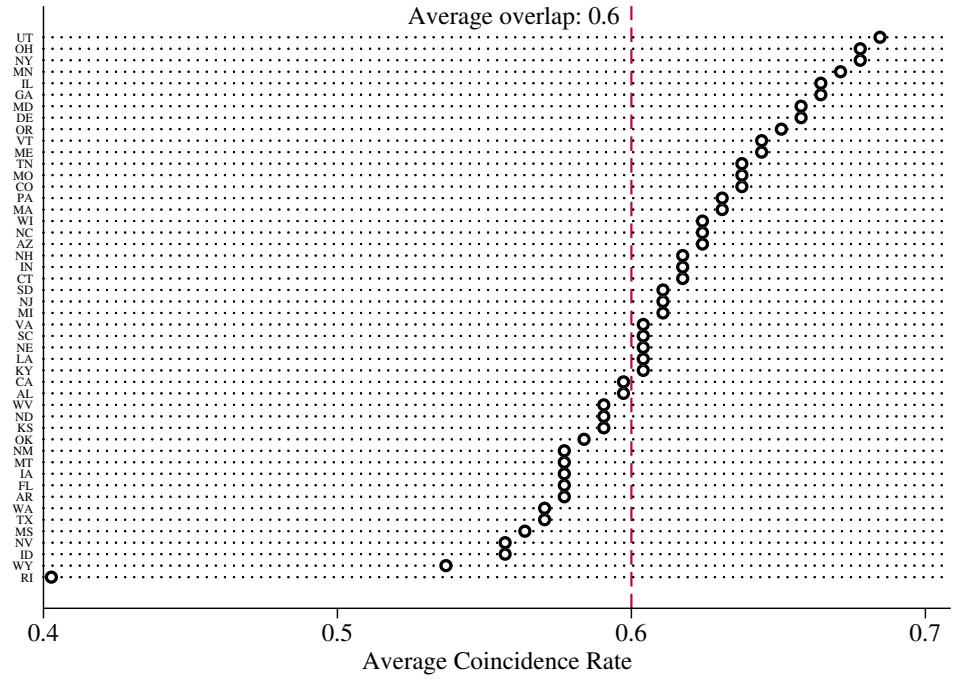
Notes: This table presents the results of bivariate panel regressions of the state-level economic activity index on several indicators not included in the baseline estimation (see Table 2). t-statistics are based on standard errors clustered by state. All indicators, except for the change in lagged sentiments, are log-transformed and standardized to have zero mean and unit variance. The change in lagged sentiments is computed as $S_{t-1} - S_{t-2}$, where S_t denotes the standardized sentiments index with zero mean and unit variance. All regressions include state fixed effects, with standard errors clustered at the state level. Statistical significance is indicated by $***$, $**$, and $*$ for the 1%, 5%, and 10% levels, respectively.

1873, 1929, and 2007. For each event, we calculate the mean change in the index for the recession years. These maps highlight that economic downturns are highly unequal in space: while most states experienced downturns, the extent to which they do varies dramatically.

To get a sense of how closely each state's economy is aligned with the U.S. business cycle, we take an approach similar to Arias, Gascon and Rapach (2016). In particular, we calculate how often a state-level recession coincides with a national one, which allows us to assess the degree of overlap between local and aggregate cycles. We measure U.S. business cycle turning points using NBER recession dates. Figure 8 shows the results. States are ordered by the degree of overlap between state and national business cycle phases, which we calculate as the fraction of times where a state and the U.S. as a whole are both signaling a recession or expansion phase. States such as Ohio or Nevada are closely aligned with the aggregate business cycle, but others such as Maine or North Dakota are not.

In sum, our analysis suggests substantial heterogeneity, both across space and time, in how much local economic cycles coincide with those of the U.S. as a whole. While a full-fledged study of long-run changes in local business cycle synchronization is beyond the scope of this paper, we believe it is worth examining in future work.

Figure 8: Estimated Average Coincidence Rate by States



Notes: We define state-level recessions with the Bry and Boschan (1971*a*) algorithm applied to our Economic Activity Index (demeaned and rescaled to levels). We define national recessions as follows: Recession years from 1887 to 1991 are defined based on Table 3 of Romer (1999), with a year counted as a recession year if it reports at least one quarter within the peak-to-trough phase. Recession years prior to 1887 are defined according to Table 1 of Davis (2006), with a year counted as a recession year if it falls within the peak-to-trough phase. For years after 1991, the NBER chronology is used.

4.5 Dating State-Level Recessions

One can use our estimates of state-level economic activity to date state recessions, similar to NBER's business cycle dating. As our analysis above highlights, business cycles vary widely across states and may differ from nationwide upswings and recessions. The principal challenge is thus to identify which periods we should classify as a state-level recession. As an illustrative method, we use the turning point algorithm first proposed by Bry and Boschan (1971*a*), a workhorse method for identifying recessions (e.g., Davis, 2006).

To implement the Bry-Boschan algorithm, we use our state-level economic indices, demeaned and in levels, and use the algorithm to identify peaks and troughs. This requires us to specify three parameters: the time window over which to identify turning points, the minimum length of expansions or contractions, and the overall duration of the cycle. Given that we have annual data, we choose a time window of two years, a minimum of one year for the length of each phase of the cycle, and an overall cycle length of two years.

Figure C.6 plots several examples of the identified peaks and troughs for California and Massachusetts, against U.S.-wide recession events in the background. We define U.S. recessions in three steps. Recession years prior to 1887 are defined according to Table 1 of Davis (2006). From 1887 to 1991, they are defined based on Table 3 of Romer (1999). In both cases, a year counts as a recession if at least one quarter (or the whole year) falls within the peak-to-trough phase. After 1991, we use the NBER to identify recession dates. In these case studies, state-level recessions tend to coincide with national recessions, but there are also exceptions. For example, California experienced a recession in 1985 that did not coincide with the dating in Romer (1999). Put differently, local and U.S.-wide recessions are clearly correlated but distinct events.

Taken together, our new chronology of state-level recession dates again highlights the considerable heterogeneity in business cycles across regions. This simple "0 or 1" measure has the potential to be a simplified indicator for local booms and busts. It is worth noting that while the Bry-Boschan method is straightforward to implement and simple to interpret, one cannot use it for a real-time identification of recessions because it requires information about future values to determine whether any given data point should be considered a turning point. Since our paper is not concerned with forecasting, we leave the application of more sophisticated methods such as Markov regime-switching models for future work.

5 Conclusion

We introduce a new historical state-level dataset for the United States, covering 60 variables from the Civil War until today. These newly constructed time series, based on 113 unique sources and a large-scale digitization effort, allow us to gauge changes in the spatial distribution of economic activity over a long span of time. In this paper, we apply this dataset to the analysis of state-level business cycles.

We estimate an annual index of state-level economic activity based on a subset of these indicators using a mixed-frequency dynamic factor model, and show that the resulting index is a reliable measure of state business cycles. Equipped with this new index, we document several new facts about economic fluctuations at the states. First, state-level business cycles exhibit substantial heterogeneity over the long run. Second, state-level business cycle volatility has not changed much over the past 150 years apart from the Great Depression and World War II periods. Third, state cycles have become more synchronized after World War II, suggesting stronger risk sharing across

states.

We also show that state-level cycles can at times diverge quite meaningfully from national cycles, and these differences in "business cycle beta" vary across states. As a by-product of our index of state-level economic activity, we introduce an NBER-style chronology of business cycle events. Different from existing work, our dating scheme has a regional dimension. We show that many recessionary periods on the state-level do not coincide with U.S.-wide downturns, highlighting the considerable variability underlying aggregate numbers.

Our work sheds new light on the history of the U.S. economy at the regional level before the advent of state-level GDP in the 1960s. As such, we view it as a starting point for more research on the nature of economic growth and fluctuations from a regional and historical perspective, made possible by our novel dataset as well as the state-level index.

References

Arias, Maria A., Charles S. Gascon and David E. Rapach. 2016. "Metro Business Cycles." *Journal of Urban Economics* 94:90–108.

Aruoba, S. B., F. X. Diebold and C. Scotti. 2009. "Real-Time Measurement of Business Conditions." *Journal of Business & Economic Statistics* 27(4):417–427.

Balke, Nathan S and Robert J Gordon. 1989. "The Estimation of Prewar Gross National Product: Methodology and New Evidence." *Journal of Political Economy* 97(1):38–92.

Barro, Robert J. and Xavier Sala-i-Martin. 1991. "Convergence Across States and Regions." *Brookings Papers on Economic Activity* 22(1):107–182.

Barro, Robert J. and Xavier Sala-i-Martin. 1992. "Convergence." *Journal of Political Economy* 100(2):223–251.

Baumeister, C., D. Leiva-León and E. Sims. 2024. "Tracking Weekly State-Level Economic Conditions." *The Review of Economics and Statistics* 106:483–504.

Berkes, Enrico. 2018. "Comprehensive Universe of US Patents (CUSP): Data and Facts." *Working Paper* .

Blanchard, Olivier and Lawrence Katz. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity* 23(1):1–76.

Bokun, Kathryn O., Laura E. Jackson, Kevin L. Kliesen and Michael T. Owyang. 2023. "FRED-SD: A Real-Time Database for State-Level Data with Forecasting Applications." *International Journal of Forecasting* 39(1):279–297.

Bry, Gerhard and Charlotte Boschan. 1971*a*. *Cyclical Analysis of Time Series: Selected Procedures and Computer Programs*. National Bureau of Economic Research.

Bry, Gerhard and Charlotte Boschan. 1971*b*. Programmed Selection of Cyclical Turning Points. In *Cyclical analysis of time series: Selected procedures and computer programs*. NBER pp. 7–63.

Burns, Arthur F and Wesley C Mitchell. 1946. *Measuring Business Cycles*. National Bureau of Economic Research.

Cattaneo, Matias D, Richard K Crump, Max H Farrell and Yingjie Feng. 2024. "On Binscatter." *American Economic Review* 114(5):1488–1514.

Clayton-Matthews, Alan and James H Stock. 1998. "An Application of the Stock/Watson Index Methodology to the Massachusetts Economy." *Journal of Economic and Social Measurement* 25(3-4):183–233.

Crone, T. M. and A. Clayton-Matthews. 2005. "Consistent Economic Indexes for the 50 States." *The Review of Economics and Statistics* 87(4):593–603.

Davis, Joseph H. 2004. "An Annual Index of U. S. Industrial Production, 1790–1915." *The Quarterly Journal of Economics* 119(4):1177–1215.

Davis, Joseph H. 2006. "An Improved Annual Chronology of US Business Cycles Since the 1790s." *The Journal of Economic History* 66(1):103–121.

DeLong, J. Bradford and Lawrence H. Summers. 1986. The Changing Cyclical Variability of Economic Activity in the United States. In *The American Business Cycle: Continuity and Change*. NBER Chapters National Bureau of Economic Research, Inc pp. 679–734.

Denton, Frank T. 1971. "Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization." *Journal of the American Statistical Association* 66(333):99–102.

Donaldson, Dave and Richard Hornbeck. 2016. "Railroads and American Economic Growth: A "Market Access" Approach." *The Quarterly Journal of Economics* 131(2):799–858.

Frickey, Edwin. 1947. *Production in the United States, 1860–1914*. Cambridge, MA: Harvard University Press.

Hamilton, James D. and Michael T. Owyang. 2012. "The Propagation of Regional Recessions." *The Review of Economics and Statistics* 94(4):935–947.

Hansen, Mary Eschelbach, Matthew Davis and Megan Fasules. 2016. United States Bankruptcy Statistics by District. 1899-2007. Technical report Inter-University Consortium for Political and Social Research.

Hindman, Monty. 2010. The Rise and Fall of Wealth Taxation: An Inquiry Into the Fiscal History of the American States PhD thesis University of Michigan.

Hodrick, Robert J. and Edward C. Prescott. 1997. "Postwar U.S. Business Cycles: An Empirical Investigation." *Journal of Money, Credit and Banking* 29(1):1–16.

Hoon, Joseph, Chang Liu, Jonathan Payne, Karsten Müller and Zhongxi Zheng. 2025. "The Costs of Financial Crises in the United States." *Working Paper* .

Kalemli-Özcan, Sebnem, Elias Papaioannou and José-Luis Peydró. 2013. "Financial Regulation, Financial Globalization, and the Synchronization of Economic Activity." *The Journal of Finance* 68(3):1179–1228.

Lewis, D. J., K. Mertens, J. H. Stock and M. Trivedi. 2022. "Measuring Real Activity Using a Weekly Economic Index." *Journal of Applied Econometrics* 37(4):667–687.

Long, John B. and Charles I. Plosser. 1983. "Real Business Cycles." *Journal of Political Economy* 91(1):39–69.

McConnell, Margaret M. and Gabriel Perez-Quiros. 2000. "Output Fluctuations in the United States: What Has Changed since the Early 1980's?" *American Economic Review* 90(5):1464–1476.

Miron, Jeffrey A. and Christina D. Romer. 1990. "A New Monthly Index of Industrial Production, 1884-1940." *The Journal of Economic History* 50(2):321–337.

Nakamura, Emi and Jón Steinsson. 2018. "Identification in Macroeconomics." *Journal of Economic Perspectives* 32(3):59–86.

Orchard, Jacob, Valerie A Ramey and Johannes F Wieland. 2024. "Micro MPCs and Macro Counterfactuals: The Case of the 2008 Rebates." *Working Paper* .

Owyang, Michael T., David E. Rapach and Howard J. Wall. 2009. "States and the Business Cycle." *Journal of Urban Economics* 65(2):181–194.

Owyang, Michael T., Jeremy Piger and Howard J. Wall. 2005. "Business Cycle Phases in U.S. States." *The Review of Economics and Statistics* 87(4):604–616.

Romer, Christina. 1986. "The Instability of the Prewar Economy Reconsidered: A Critical Examination of Historical Macroeconomic Data." *The Journal of Economic History* 46(2):494–496.

Romer, Christina D. 1989. "The Prewar Business Cycle Reconsidered: New Estimates of Gross National Product, 1869-1908." *Journal of Political Economy* 97(1):1–37.

Romer, Christina D. 1999. "Changes in Business Cycles: Evidence and Explanations." *Journal of Economic perspectives* 13(2):23–44.

Simpson, P. B. and P. S. Anderson. 1957. "Liabilities of Business Failures as a Business Indicator." *The Review of Economics and Statistics* 39(2):193–199.

Stock, J. H. and M. W. Watson. 1989. "New Indexes of Coincident and Leading Economic Indicators." *NBER Macroeconomics Annual* 4:351–394.

Stock, J. H. and M. W. Watson. 1991. *A Probability Model of the Coincident Economic Indicators*. Cambridge University Press p. 63–90.

Stock, James H and Mark W Watson. 1999. "Business Cycle Fluctuations in Us Macroeconomic Time Series." *Handbook of Macroeconomics* 1:3–64.

Stock, James H. and Mark W. Watson. 2002. "Has the Business Cycle Changed and Why?" *NBER Macroeconomics Annual* 17:159–218.

Sylla, Richard E., John B. Legler and John Wallis. 1993. "Sources and Uses of Funds in State and Local Governments, 1790-1915: [United States].".

Van Binsbergen, Jules H, Svetlana Bryzgalova, Mayukh Mukhopadhyay and Varun Sharma. 2024. (Almost) 200 Years of News-Based Economic Sentiment. Technical report National Bureau of Economic Research.

Wallis, John Joseph. 2000. "American Government Finance in the Long Run: 1790 to 1990." *Journal of Economic Perspectives* 14(1):61–82.

Williamson, S. H. 2025. "What was the U.S. GDP then?" MeasuringWorth, 2025.

Online Appendix for "U.S. State-Level Business Cycles Since the Civil War"

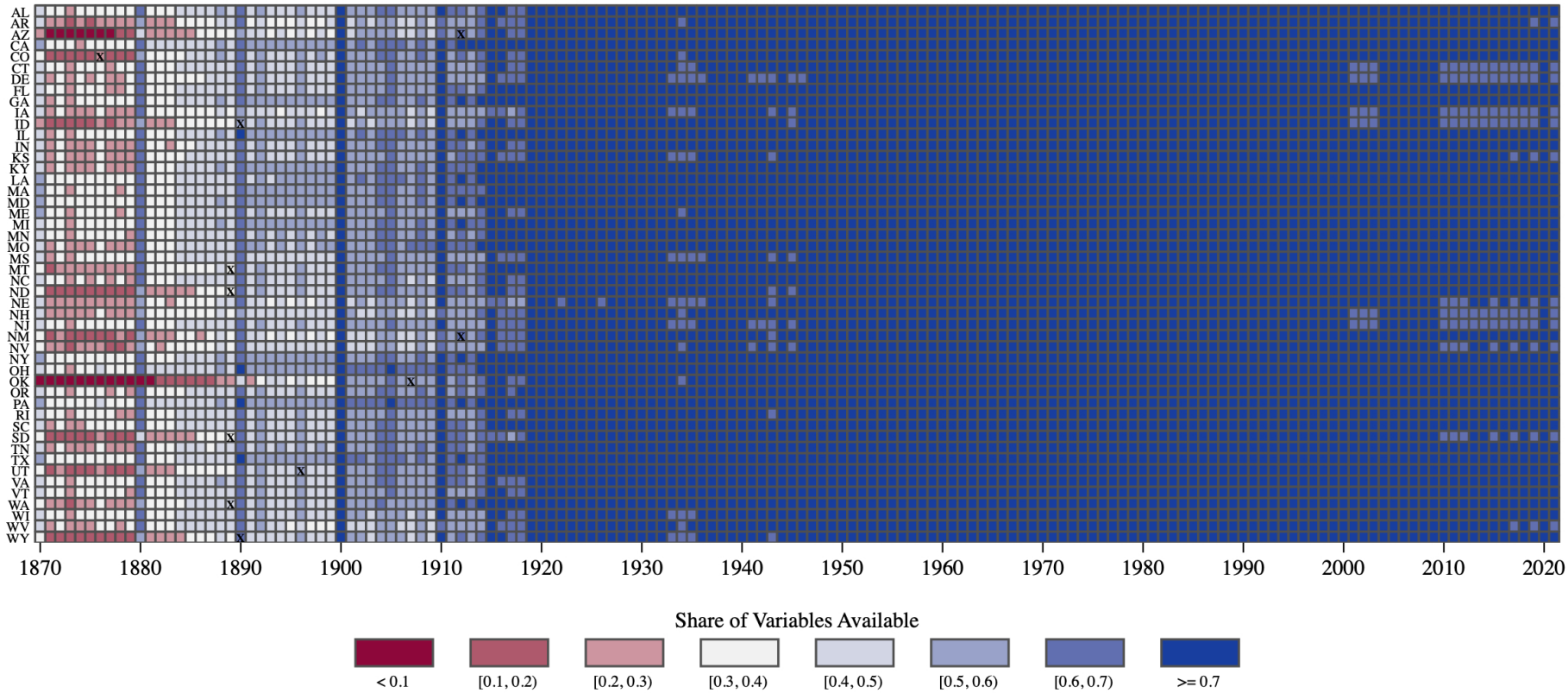
Joseph Hoon Chang Liu Karsten Müller Zhongxi Zheng

July 6, 2025

A Details on the Dataset

Figure A.1 displays the fraction of available input variables for each state-year observation. Data availability is more limited in the earlier years. Nonetheless, the dataset exhibits relatively strong coverage in core sectors—namely agriculture, mining, and manufacturing—which collectively represent the bulk of economic activity during that period. Additional details on data construction, variable definitions, and source documentation are provided in the supplementary appendix by Hoon, Liu, Müller and Zheng (2025).

Figure A.1: Variable Coverage by State



Notes: This figure shows the share of variables in the dataset that are available in a given year for the 48 contiguous states. We plot black crosses to indicate the year of a state's admission to the Union. We observe that for the majority of states, including but not limited to Arizona (admitted in 1912) and Oklahoma (admitted in 1907), variable coverage improves post-admission.

B The Gibbs Sampling Algorithm

Let $\boldsymbol{\theta}$ collect the model parameters in the state space system (5)–(6). Conditional on $\boldsymbol{\theta}$ and $\mathcal{F}_T$, the first step involves drawing $\boldsymbol{\alpha}_t$ using the Kalman filter and smoothing recursions; see Carter and Kohn (1994) and Durbin and Koopman (2012) for a detailed treatment of the Kalman filter and smoothing.[7] It is likely that we do not observe all indicators in a given year t. In this case, we remove the rows of $\mathbf{H}_t$ that are associated with the missing entries. This operation ensures that $\mathbf{H}_t$ is conformable in the observation equation when we perform the first step. In the second step, we take the draws of $\boldsymbol{\alpha}_t$ as given, and proceed to update $\boldsymbol{\theta}$ based on Bayesian methods. In particular, we follow Baumeister, Leiva-León and Sims (2024) to assume that the elements of $\boldsymbol{\theta}$ are distributed by natural-conjugate priors; and therefore, the property of conjugacy ensures that the posterior distribution belongs to the same class of probability distribution as the priors. We assume that $\{\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\lambda}\}$ have Gaussian priors with the typical setup of zero mean and unit variances. Given the state equation and the assumption that $\boldsymbol{\psi}$ has a Gaussian prior, a natural-conjugate prior for $\boldsymbol{\sigma}$ is the inverse Gamma distribution. In our baseline specification, we assume that the first two parameters have values of 10 and 0.9, respectively. We refer readers to Gelman et al. (2013) for details of the Gibbs sampler.

C Additional Figures and Tables

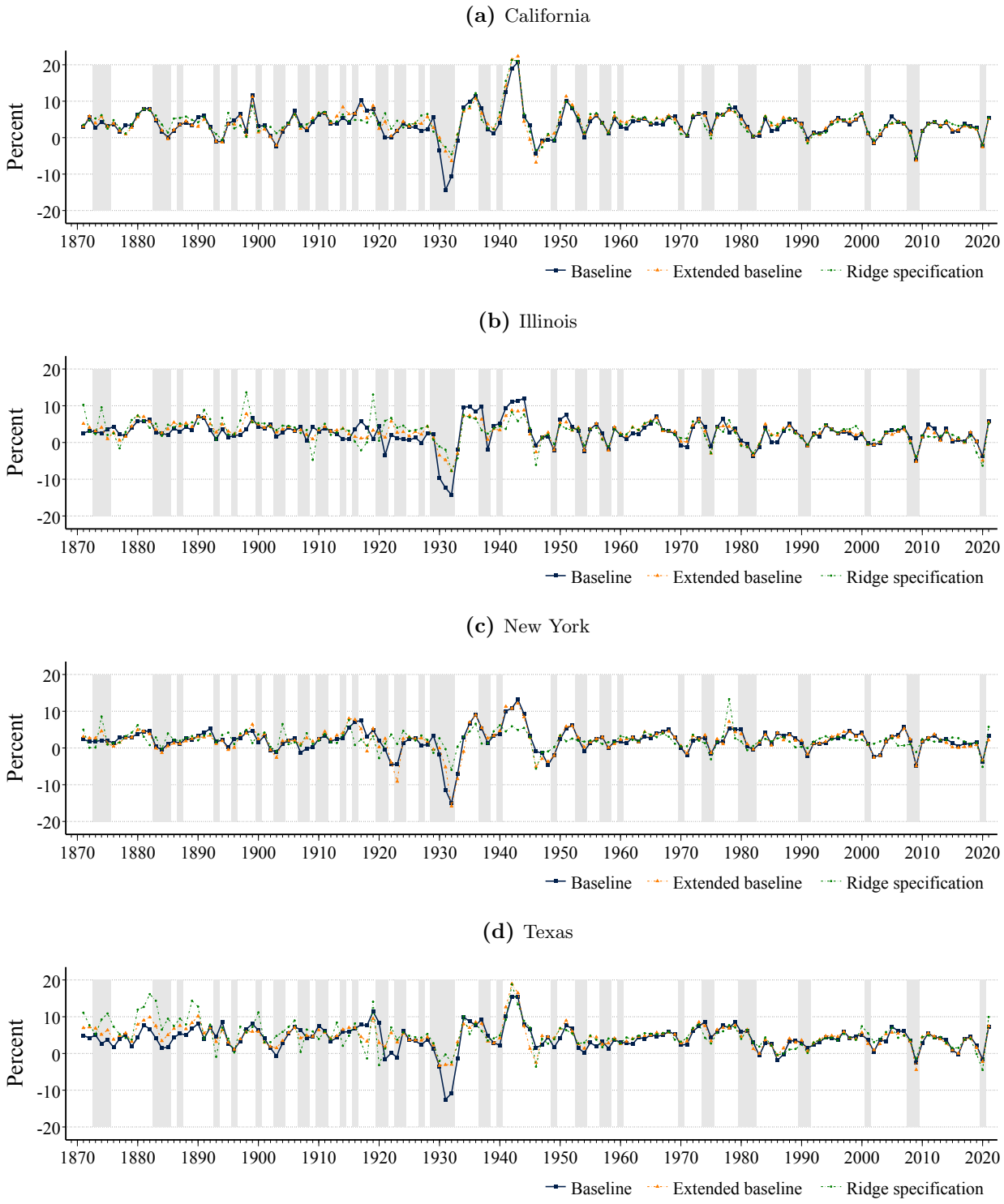
[7] We have assumed that $\boldsymbol{\eta}_t$ follows a normal distribution in the state equation. We note in passing that the Gaussian assumption is not necessary to use the Kalman filter recursion; and in fact, if the Gaussian assumption is not correct, the estimates of $\boldsymbol{\theta}$ are still consistent, albeit not efficient.

Table C.1: Descriptive Statistics of the Estimated SEAI

State	Average	Volatility	25th Percentile	Median	75th Percentile
Alabama	3.35	4.41	1.78	3.50	5.39
Arkansas	4.00	4.35	2.34	3.98	6.20
Arizona	5.38	5.31	3.17	5.21	8.34
California	3.70	3.97	1.83	3.78	5.65
Colorado	4.52	4.50	2.48	4.72	6.17
Connecticut	2.75	4.18	0.92	2.75	4.92
Delaware	3.11	4.82	0.31	2.95	5.71
Florida	4.64	4.29	2.53	4.75	7.07
Georgia	4.36	4.67	2.48	4.86	6.61
Iowa	3.09	4.89	1.25	3.72	5.44
Idaho	4.48	5.34	1.98	4.57	7.26
Illinois	2.58	3.57	0.95	2.82	4.21
Indiana	3.08	5.17	0.41	2.99	6.23
Kansas	2.91	2.98	1.61	2.91	4.55
Kentucky	3.18	5.34	0.45	3.47	5.12
Louisiana	2.33	4.95	0.28	2.60	4.92
Massachusetts	2.83	3.11	1.64	3.11	4.37
Maryland	3.52	3.75	1.80	3.47	5.12
Maine	1.99	3.39	0.30	1.59	3.64
Michigan	2.49	7.04	−0.37	1.93	5.71
Minnesota	3.89	3.94	2.24	4.14	5.86
Missouri	2.50	3.72	0.85	2.70	4.32
Mississippi	3.47	6.28	0.65	3.16	6.30
Montana	3.09	4.91	0.87	3.18	6.06
North Carolina	4.43	4.03	2.89	4.38	6.59
North Dakota	4.52	10.37	0.41	5.36	11.15
Nebraska	3.60	4.64	1.47	3.29	5.98
New Hampshire	3.31	3.63	1.78	3.27	5.53
New Jersey	3.11	3.84	1.61	3.12	5.10
New Mexico	4.11	5.87	1.94	3.88	6.64
Nevada	4.36	5.20	2.01	4.16	6.56
New York	2.03	3.33	1.09	2.33	3.34
Ohio	2.28	4.59	0.19	2.46	4.30
Oklahoma	3.06	4.15	1.62	3.33	5.02
Oregon	4.07	4.58	1.49	4.27	6.85
Pennsylvania	2.23	3.37	1.03	2.30	3.80
Rhode Island	2.23	3.86	0.60	2.50	4.24
South Carolina	4.46	4.21	2.51	4.68	6.92
South Dakota	5.49	7.28	1.26	4.39	10.33
Tennessee	4.12	5.06	2.15	4.38	6.73
Texas	4.15	3.50	2.53	4.21	6.10
Utah	4.31	3.94	2.65	4.36	6.42
Virginia	4.27	7.08	1.11	3.90	6.70
Vermont	2.48	4.57	0.73	2.72	4.38
Washington	4.59	5.59	1.94	4.50	7.23
Wisconsin	2.93	3.49	1.62	3.23	4.26
West Virginia	2.20	3.54	0.76	2.39	4.49
Wyoming	2.91	5.11	1.16	3.01	5.77

Notes: This table presents the descriptive statistics for the state-level economic activity indices from 1871 to 2021.

Figure C.2: Alternative Estimates of Economic Activity Indices for Selected States



Notes: In this figure, *Baseline* refers to the estimation using input indicators listed in Table 2, while *Extended baseline* adds (i) total bank assets and liabilities from Hoon, Liu, Payne, Müller and Zheng (2025) and (ii) changes in lagged sentiments from Van Binsbergen et al. (2024). See the notes in Table 4 for details on constructing changes in lagged sentiments. *Ridge specification* refers to the estimation using input indicators selected via ridge regression, where, for each state, state-level personal income is regressed on a pool of indicators, including those in the *Extended baseline* and Table 4. The procedure is as follows: (i) pick 1000 logarithmically spaced ridge parameters between 10^{-4} and 10^{4}; (ii) regress state-level personal income on this pool of indicators for each ridge parameter and obtain the regression slope coefficients; (iii) compute the absolute average of the coefficients for each indicator; (iv) include an indicator if its coefficient exceeds the 30th percentile of all coefficients.

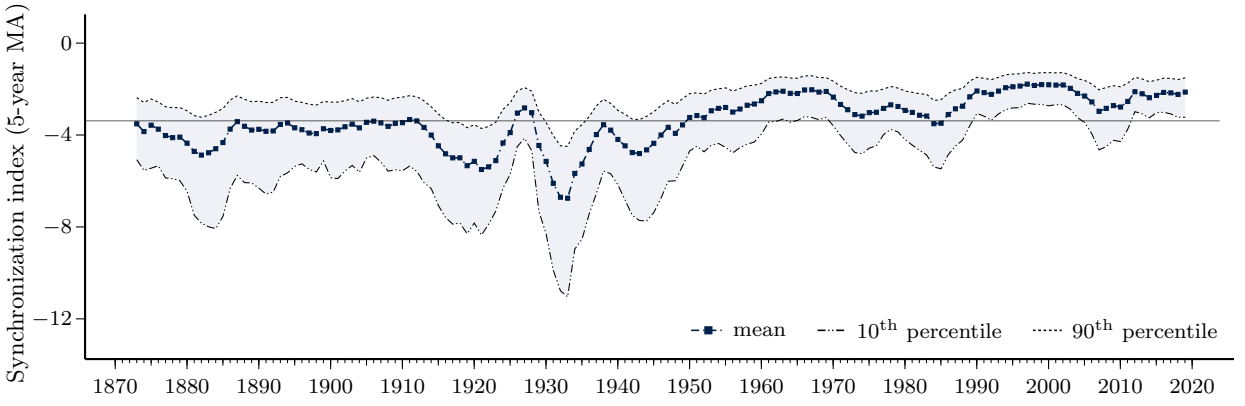
C.1 Synchronization

For our baseline estimation of business cycle synchronization, we calculate the cross-state standard deviation of changes in economic activity in each year. In this section, we use an alternative approach following Kalemli-Özcan, Papaioannou and Peydró (2013). In particular, we calculate a synchronization measure for state i as the sum of negative absolute differences between the state's economic activity index and those of all other states in a given year, scaled by the total number of state pairs:

$$Synchronization_{i,t} = -\frac{\sum_{i \neq i'} |s_{i,t} - s_{i',t}|}{S_t - 1}, \tag{C.1}$$

where $s_{i,t}$ and $s_{i',t}$ refer to the scaled economic activity index, according to equation (8), for states i and i' in year t. $S_t \leq 48$ denotes the number of states for which the scaled economic activity index is available in year t. From equation (C.1), state i's economic activity is more synchronized with those of the other states as $Synchronization_{i,t}$ approaches zero. In Figure C.3, we report the mean, 10th, and 90th percentiles of $Synchronization_{i,t}$ over time. From this figure, we see that the economic activity indices seem to exhibit a mild U-shape pattern, where the states' business cycles are least synchronized during the 1930s–40s, and they are increasingly more synchronized after the 1950s.

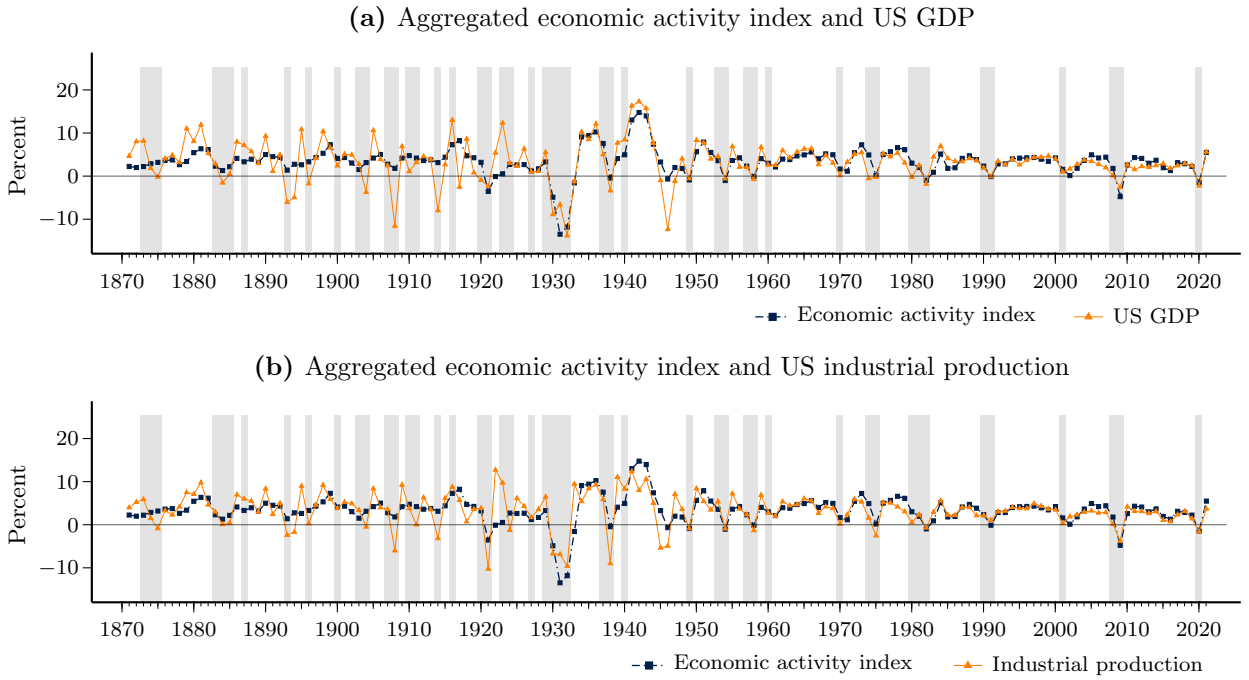
Figure C.3: Synchronization of state-level economic activity over time



Notes: This figure presents the mean, 10th, and 90th percentiles of the synchronization index, averaged over a five-year moving window. The synchronization index is computed using Equation (C.1), where the number of states, S_t, ranges from 41 to 48, depending on the availability of economic activity indices shown in Figure 4. The solid line denotes the average of the mean index level over time, which is approximately −3.4.

C.2 Aggregated State-Level Index vs U.S. Measures

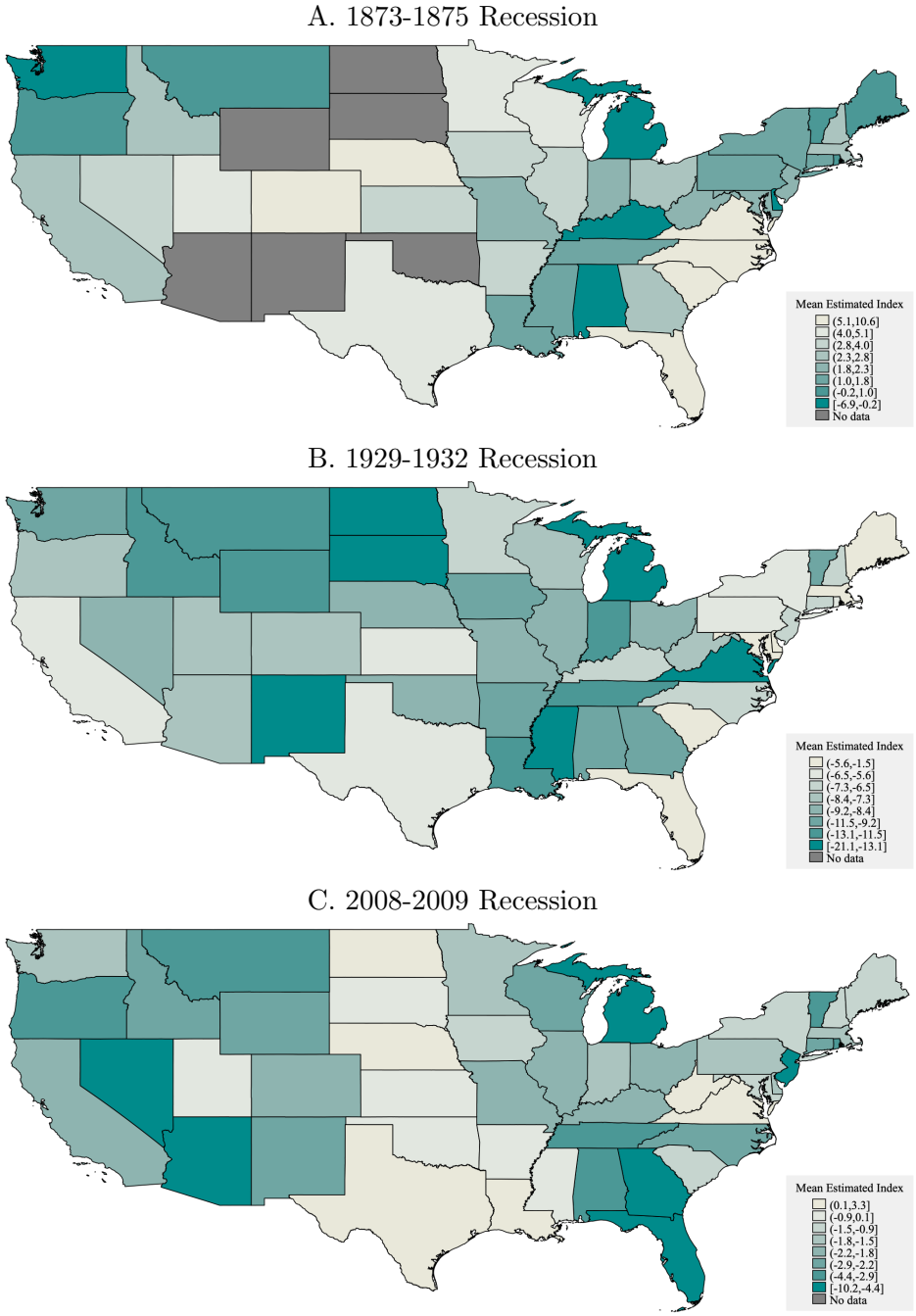
Figure C.4: Comparison of the Aggregated SEAI and Other US-Wide Measures



Notes: This figure plots the aggregated economic activity index alongside US GDP and industrial production from 1871 to 2019. The aggregated index is constructed by taking a weighted average of the state-level economic activity indices, with the weights based on the relative size of each state's economy compared to the sum across all 48 states. For each state, economic size is measured by the level of its economic activity index, scaled so that the 2012 value matches the state's GDP in 2012 dollars. The industrial production series is constructed by combining the data from Davis (2004) (1871–1915), Miron and Romer (1990) (1916–1919), and those published by the Fed (1920–2019). Both the aggregated index and industrial production are scaled and retrended to US GDP. The US GDP data are sourced from Williamson (2025). The shaded bars indicate recession years; see the notes to Figure 6 for their definition.

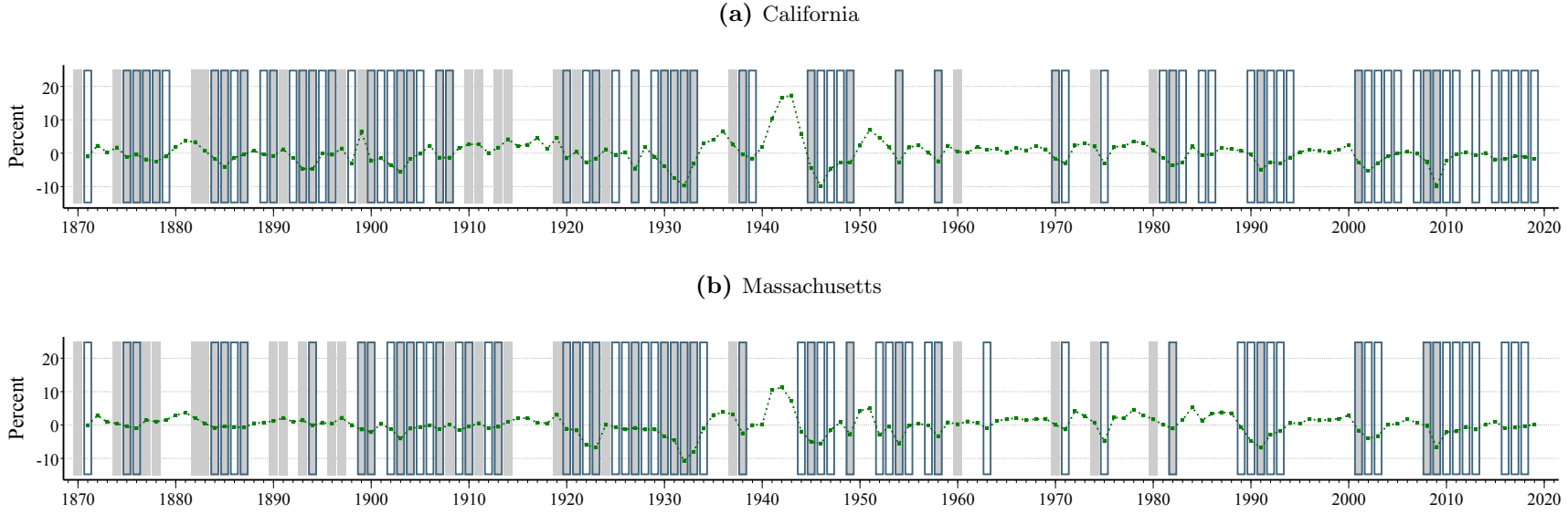
C.3 State-Level Economic Activity During US-Wide Recessions

Figure C.5: Changes in State-Level Economic Activity During US-Wide Recessions



Notes: This figure shows the average percentage changes in economic activity indices during three national recession episodes: the 1873–75 Recession, the Great Depression (1929–32), and the Great Recession (2008-09). See the notes in Figure 6 for the definition of national recessions.

Figure C.6: Recession dates for selected states (1871–2019)



Notes: Recession dates for the states are identified by applying the Bry and Boschan algorithm (1971) to the economic condition indices (demeaned and scaled to levels). The gray bars correspond to the NBER recession dates and the dashed lines represent the demeaned economic condition indices.

References

Baumeister, C., D. Leiva-León and E. Sims. 2024. "Tracking Weekly State-Level Economic Conditions." *The Review of Economics and Statistics* 106:483–504.

Bry, Gerhard and Charlotte Boschan. 1971. *Cyclical Analysis of Time Series: Selected Procedures and Computer Programs*. National Bureau of Economic Research.

Carter, C. K. and R. Kohn. 1994. "On Gibbs Sampling for State Space Models." *Biometrika* 81(3):541–553.

Davis, Joseph H. 2004. "An Annual Index of U. S. Industrial Production, 1790–1915." *The Quarterly Journal of Economics* 119(4):1177–1215.

Durbin, J. and S. J. Koopman. 2012. *Time Series Analysis by State Space Methods*. Oxford University Press.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin. 2013. *Bayesian Data Analysis*. Chapman and Hall/CRC.

Hoon, Joseph, Chang Liu, Jonathan Payne, Karsten Müller and Zhongxi Zheng. 2025. "The Costs of Financial Crises in the United States." *Working Paper* .

Hoon, Joseph, Chang Liu, Karsten Müller and Zhongxi Zheng. 2025. "A U.S. State-Level Dataset: 1863-2021." *Working Paper* .

Kalemli-Özcan, Sebnem, Elias Papaioannou and José-Luis Peydró. 2013. "Financial Regulation, Financial Globalization, and the Synchronization of Economic Activity." *The Journal of Finance* 68(3):1179–1228.

Miron, Jeffrey A. and Christina D. Romer. 1990. "A New Monthly Index of Industrial Production, 1884-1940." *The Journal of Economic History* 50(2):321–337.

Van Binsbergen, Jules H, Svetlana Bryzgalova, Mayukh Mukhopadhyay and Varun Sharma. 2024. (Almost) 200 Years of News-Based Economic Sentiment. Technical report National Bureau of Economic Research.

Williamson, S. H. 2025. "What was the U.S. GDP then?" MeasuringWorth, 2025.

A U.S. State-Level Dataset: 1863-2021*

Joseph Hoon† Chang Liu‡ Karsten Müller§ Zhongxi Zheng¶

July 6, 2025

Abstract

We document the details of a new comprehensive U.S. state-level dataset introduced in Hoon, Liu, Müller and Zheng (2025) covering the period from 1863 to today, including the data sources, imputation methods, temporal disaggregation methods, and many additional nuances involved in constructing historical time series from many different sources. We also provide additional statistics on the availability of each variable in our dataset for each state.

Keywords: state-level data; economic history

JEL classification: N91, N92, R10

*This paper provides details to the state-level dataset used in "U.S. State-Level Business Cycles Since the Civil War" by the same authors. A complete list of acknowledgments is provided in Section 5.

†Department of Economics, National University of Singapore. Email: joseph.hoon@u.nus.edu.

‡Department of Economics and Risk Management Institute, National University of Singapore. Email: charlesliu.pku@gmail.com.

§Department of Finance and Risk Management Institute, National University of Singapore. Email: kmueller@nus.edu.sg.

¶Department of Economics, National University of Singapore. Email: zhongxi.zheng@u.nus.edu.

CONTENT

1 An Overview of the Dataset

Hoon, Liu, Müller and Zheng (2025) introduce a historical state-level dataset covering 60 variables for 48 states over the period 1863–2021, constructed from a grand total of 113 sources. Table 2 presents an overview of the variables that make up our raw data, including detailed information on the main sources we use to compile them, the variables collected from each source, and their coverage across states and time.

2 Methodology

2.1 Data Digitization

Figure 1 provides an example of the sources from which we digitize our data and the corresponding variables extracted. We follow the following procedure to digitize data: first, we run the primary sources through Optical Character Recognition Software (AWS Textract). Second, we manually verify the accuracy of the data. Often, regional and national totals are reported along with the state-level data, as is the case in Figure 1b. Whenever this is the case, we run verification checks by ensuring that the sum of the data reported for individual states tallies with the regional totals. This internal reconciliation serves as an additional layer of quality control, allowing us to identify issues that might not be evident through row-by-row verification.

2.2 Imputation of Annual Sales Receipts for Agricultural Variables

This section outlines the procedure for imputing annual sales receipts of livestock, crops, and forest products for the 48 contiguous U.S. states from 1870 to 1924. To begin, let $Y_{i,t}^{l}$, $Y_{i,t}^{c}$, and $Y_{i,t}^{f}$ denote the sales receipts from livestock (l), crops (c), and forest products (f) for state i in year t. For each $j \in \{l, c, f\}$, let $\mathcal{T}_i^j$ denote the set of years in which $Y_{i,t}^{j}$ is observed within the imputation period;[1] here, to simplify exposition, we assume $\mathcal{T}_i^j = \mathcal{T} \equiv \{1870, 1880, 1890, 1900, 1910, 1919, 1924\}$ for all states i and categories j.[2] Let $\overline{\mathcal{T}}$ denote the set of all years in the imputation period excluding the first year (i.e., 1870). Then, the set of years for which we impute values of $Y_{i,t}^{j}$ is given by $\overline{\mathcal{T}} \backslash \mathcal{T}$.

[1] The imputation period is defined as the interval spanning from the earliest year in which $Y_{i,t}^{j}$ is observed to 1924.

[2] From our dataset, we have $\mathcal{T}_i^l = \mathcal{T}$ for all but one state, and $\mathcal{T}_i^c = \mathcal{T}$ for 36 states. For any state i' with $\mathcal{T}_{i'}^j \neq \mathcal{T}$, the imputation procedure described here remains valid, provided that the entries in $\mathcal{T}$ and $\overline{\mathcal{T}}$ are adjusted accordingly.

To impute the missing observations for $Y^l_{i,t}$ and $Y^c_{i,t}$, we rely on the fact that our dataset contains annual sales receipts for major livestock and crop commodities over much of the imputation period, although some individual series become available only after 1870.[3] Mathematically, let $X^l_{i,t} = [X^l_{i,t,1}, \ldots, X^l_{i,t,n}]$ denote the vector of sales receipts for n major livestock commodities, and similarly, let $X^c_{i,t} = [X^c_{i,t,1}, \ldots, X^c_{i,t,m}]$ represent the vector of sales receipts for m major crop commodities. Now, let $x^l_{i,t} = \sum_s X^l_{i,t,s}$ and $x^c_{i,t} = \sum_s X^c_{i,t,s}$ denote the total sales receipts from these livestock and crop commodities, respectively. Table 1 outlines the commodities used in the imputation—see Section 3.1 for further documentation of these individual series in our dataset. Since $x^j_{i,t}$ likely constitutes a large portion of $Y^j_{i,t}$, we postulate a strong positive correlation between their year-on-year growth rates for each $j \in \{l, c\}$. Let $\tilde{x}^j_{i,t} = \ln(x^j_{i,t}/x^j_{i,t-1})$ and $\tilde{y}^j_{i,t} = \ln(Y^j_{i,t}/Y^j_{i,t-1})$ denote the respective log growth rates. We then determine the values of $\tilde{y}^j_{i,t}$ by solving a constrained minimization problem à la Denton (1971). Formally, for each state i and category j, we solve:

$$\min_{\{\tilde{y}^j_{i,t}\}} \sum_{t \in \overline{\mathcal{T}}} \left(\Delta^h \tilde{y}^j_{i,t} - \Delta^h \tilde{x}^j_{i,t}\right)^2 \quad \text{s.t.} \quad Y^j_{i,t} = \hat{Y}^j_{i,t} \ \text{ for all } \ t \in \mathcal{T}^j_i = \mathcal{T}, \tag{1}$$

where Δ^h denotes the hth-order difference operator, and $\hat{Y}^j_{i,t}$ denotes the observed value of $Y^j_{i,t}$. For a given initial value $\hat{Y}^j_{i,1870}$, the constraint $Y^j_{i,t} = \hat{Y}^j_{i,t}$ for $t \in \mathcal{T}$ is equivalent to imposing

$$\sum_{s=t_{k-1}+1}^{t_k} \tilde{y}^j_{i,s} = \ln(\hat{Y}^j_{i,t_k}/\hat{Y}^j_{i,t_{k-1}}) \quad \text{for} \quad k = 2, \ldots, 7, \tag{2}$$

where t_k is the kth element in $\mathcal{T}$, i.e., $\mathcal{T} = \{t_1, t_2, \ldots, t_7\} = \{1870, 1879, \ldots, 1924\}$. In other words, once $Y^j_{i,1870}$ is fixed, imposing that the cumulative log growth over each intercensal block $(t_{k-1}, t_k]$ matches the observed log change ensures that the imputed series of $Y^j_{i,t}$ exactly coincides with $\hat{Y}^j_{i,t}$ at every $t \in \mathcal{T}$. In equation (1), we follow the recommendation of Boot, Feibes and Lisman (1967, Section 4) and set $h = 2$.[4] Given (1) and (2), the sequence $\{\tilde{y}^j_{i,t}\}_{t \in \overline{\mathcal{T}}}$ is then solved using the formula

[3] More precisely, sales receipts for at least three major livestock and three major crop commodities are available in any

[4] We also experiment with $h = 1$; the imputed values are qualitatively similar to those obtained here for most states.

specified in Denton (1971); that is:

$$\tilde{y}_i^j = \tilde{x}_i^j + A^{-1}B\big(B^\top A^{-1}B\big)^{-1}\big(\hat{y}_i^j - B^\top \tilde{x}_i^j\big), \quad \text{where} \quad \tilde{y}_i^j = \begin{bmatrix} \tilde{y}_{i,1871}^j \\ \vdots \\ \tilde{y}_{i,1924}^j \end{bmatrix} \in \mathbb{R}^{54}, \quad \tilde{x}_i^j = \begin{bmatrix} \tilde{x}_{i,1871}^j \\ \vdots \\ \tilde{x}_{i,1924}^j \end{bmatrix} \in \mathbb{R}^{54},$$

and $\hat{y}_i^j = \big[\ln(\hat{Y}_{i,t_2}^j/\hat{Y}_{i,t_1}^j), \ldots, \ln(\hat{Y}_{i,t_k}^j/\hat{Y}_{i,t_{k-1}}^j), \ldots, \ln(\hat{Y}_{i,t_7}^j/\hat{Y}_{i,t_6}^j)\big]^\top \in \mathbb{R}^6$. The matrix $A \in \mathbb{R}^{54\times 54}$ is defined so that the objective function in (1) takes the form $(\tilde{y}_i^j - \tilde{x}_i^j)^\top A(\tilde{y}_i^j - \tilde{x}_i^j)$, while the matrix $B \in \mathbb{R}^{54\times 6}$ is constructed to impose the constraint in (2). Finally, since $h = 2$, we note that matrix A satisfies $A^{-1} = R^\top\big(R^\top R\big)R$, where R is an upper-triangular matrix with all elements on or above the main diagonal equal to 1 and all others equal to 0.

Table 1: Major livestock and crop commodities

	Coverage			Coverage	
Livestock	No. of states	Start year	Crop	No. of states	Start year
Mules	42	1870	Hay	1	1909
Cattles & Calves	48	1870	Cotton	12	1870
Hogs	48	1870	Cottonseed	13	1870
Sheep	48	1870	Tobacco	16	1870
Horses	48	1870	Sweet potato	22	1870
			Barley	34	1870
			Rye	34	1870
			Wheat	43	1870
			Corn	47	1870
			Oats	47	1870
			Potato	47	1870

Notes: In each 3-column block, 'No. of states' indicates the number of contiguous U.S. states in which the individual series is used for imputation; 'Start' refers to the first year for which the series is available in at least one of the states.

For each $j \in \{l, c\}$, we recover the level series of $Y_{i,t}^j$ (denoted by $\tilde{Y}_{i,t}^j$) recursively from the initial observation $\hat{Y}_{i,1870}^j$ via

$$\tilde{Y}_{i,t}^j = \hat{Y}_{i,1870}^j \cdot \exp\left(\sum_{s=1871}^{t} \tilde{y}_{i,s}^j\right) \quad \text{for all} \quad t > 1870,$$

as desired.

Since our dataset does not contain annual sales receipts for major forest products—as it does

for livestock and crops—we cannot impute the missing observations using the approach described in the previous paragraphs. Instead, we interpolate $Y^f_{i,t}$ for $t \in \overline{\mathcal{T}} \setminus \mathcal{T}$ using the monotone piecewise cubic interpolation method of Fritsch and Carlson (1980). Finally, we take the sum of the imputed series for $Y^l_{i,t}$, $Y^c_{i,t}$, and $Y^f_{i,t}$ as an estimate of the annual value of agricultural products sold for state i in year t. This aggregate state-level measure is then used as an input in our factor estimation.

Robustness Here, we implement the reduced-form method of Chow and Lin (1971) as a robustness check against the constrained minimization approach used above. To fix ideas, we describe the Chow-Lin method using $Y^l_{i,t}$; the same steps apply directly to $Y^c_{i,t}$ and are omitted for brevity. Essentially, the method posits a linear relationship between Y^l_i and X^l_i of the form:

$$Y^l_i = X^l_i\beta + U^l_i, \quad \text{where} \quad Y^l_i = \begin{bmatrix} Y^l_{i,1871} \\ \vdots \\ Y^l_{i,1924} \end{bmatrix} \in \mathbb{R}^{54}, \quad X^l_i = \begin{bmatrix} 1 & X^l_{i,1871,1} & \cdots & X^l_{i,1871,n} \\ \vdots & \vdots & & \vdots \\ 1 & X^l_{i,1924,1} & \cdots & X^l_{i,1924,n} \end{bmatrix} \in \mathbb{R}^{54\times(n+1)},$$

and U^l_i is a zero-mean random vector with covariance matrix $V^l_i = \sigma^l_i W^l_i$, where W^l_i is symmetric and positive definite. Since Y^l_i contains missing observations, the model cannot be estimated directly. We address this by pre-multiplying both sides of the equation by an aggregation matrix C, yielding:

$$CY^l_i = CX^l_i\beta + CU^l_i, \quad \text{where} \quad CY^l_i = \left[\, \hat{Y}^l_{i,1880}, \hat{Y}^l_{i,1890}, \hat{Y}^l_{i,1900}, \hat{Y}^l_{i,1910}, \hat{Y}^l_{i,1919}, \hat{Y}^l_{i,1924} \,\right]^\top,$$

and the time indexes correspond to $t \in \mathcal{T} \setminus \{1870\}$. We note in passing that the matrix C is binary and row-stochastic, with exactly one entry equal to one in each row. The covariance matrix of CU^l_i is then given by $\sigma^l_i C W^l_i C^\top \equiv \sigma^l_i w^l_i$. If w^l_i is known, the standard generalized least squares estimator of β is $\tilde{\beta} = \left[(CX^l_i)^\top (w^l_i)^{-1}(CX^l_i)\right]^{-1}(CX^l_i)^\top (w^l_i)^{-1} CY^l_i$, where the regression residual is defined as $\tilde{u}^l_i \equiv C\tilde{U}^l_i = CY^l_i - CX^l_i\tilde{\beta}$. The best linear unbiased estimator of Y^l_i is $\tilde{Y}^l_i = X^l_i\tilde{\beta} + W^l_i C^\top (w^l_i)^{-1}\tilde{u}^l_i$, which consists of two components: the first term estimates the conditional expectation of Y^l_i given X^l_i, while the second term uses prior information about the covariance structure W^l_i, together with the estimated residuals $\tilde{u}^l_i$, to recover an estimate of the disturbance vector U^l_i. Next, the standard

error of the kth element of $\tilde{\beta}$ is given by:

$$\text{SE}(\tilde{\beta}_k) = \sqrt{\left[\sigma_i^l \Big((CX_i^l)^\top (w_i^l)^{-1} (CX_i^l)\Big)^{-1}\right]_{kk}} \quad \text{where} \quad \sigma_i^l = \frac{1}{p-n} (\tilde{u}_i^l)^\top (w_i^l)^{-1} (\tilde{u}_i^l),$$

and p (with $p > n$) is the dimension of CY_i^l. Note that the results above rest on the assumption that the covariance matrix W_i^l is known. In practice, however, it is unobserved and must be estimated. Moreover, the precision of the GLS estimator may be limited when p is small. We address the two issues sequentially below.

Estimating the covariance matrix of annual residuals. Following Chow and Lin (1971), the entries of U_i^l are assumed to follow an AR(1) process. Under this assumption, the structure of W_i^l reduces to a Toeplitz matrix with elements defined by:

$$W = \frac{1}{1-\varrho^2} \begin{bmatrix} 1 & \varrho & \varrho^2 & \dots & \varrho^{N-1} \\ \varrho & 1 & \varrho & \ddots & \vdots \\ \varrho^2 & \varrho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \dots & \varrho \\ \varrho^{N-1} & \dots & \dots & \varrho & 1 \end{bmatrix},$$

where ϱ is the persistence of the AR(1) process. Observe that W_i^l is symmetric and positive definite by construction. In our implementation, we select a value of $\varrho \in (-1, 1)$ by minimizing the following criterion:

$$\min_{\varrho} \ (\tilde{u}_i^l)^\top (w_i^l)^{-1} \tilde{u}_i^l, \tag{3}$$

where $\tilde{u}_i^l$ denotes the regression residual computed at each candidate value of $\varrho \in (-1, 1)$.[5]

Bypassing the small-sample issue. Two problems arise when p is very small, as in our baseline

[5]Note that the criterion in (3) is analogous to solving a GLS problem under a positive-definite covariance matrix W_i^l — see Section 3.1 of Björck (2024). To find the minimizer $\tilde{\varrho}$, we evaluate the criterion over 1,000 equally spaced grid points in the open interval $(-1, 1)$. Here, we exclude the case $\varrho = 1$, which corresponds to a random walk process for the entries in U_i^l. This case is studied in Fernández (1981) and Litterman (1983). However, an implicit assumption of that structure is the absence of a long-run relationship between $Y_{i,t}^l$ and the regressors in $X_{i,t}^l$, which does not align with our setting.

implementation of the Chow-Lin procedure: first, the estimate of β becomes imprecise; second, the requirement that $p > n$ limits the number of regressors that can be included in X_i^l. A straightforward solution then entails increasing p by appending more observations to CY_i^l; this requires that we expand the set $\mathcal{T}$ from $\{1870, 1880, 1890, 1900, 1910, 1919, 1924\}$ to $\{1870, 1880, 1890, 1900, 1910, 1919, 1924, 1925, 1926, \ldots, \mathcal{T}_s\}$, as annual observations of $Y_{i,t}^l$ are available in our dataset from 1924 onwards. In our baseline configuration, we set $\mathcal{T}_s = 1950$ for all states i and categories $j \in \{l, c\}$. To illustrate the robustness of the imputation results, Figures 2 and 3 display the imputed values of $Y_{i,t}^c$ and $Y_{i,t}^l$, respectively, for eight individual states from 1970 to 1923, using both the baseline and Chow-Lin methods. The results show that both approaches yield qualitatively similar estimates across the selected states.

3 Details on the Dataset

We organize the state-level data into eight categories: agriculture, manufacturing, mining, transportation, business statistics, government finances, labor market outcomes, and other variables, which we outline in more detail in Sections 3.1 to 3.8. In contrast to the main text, we organize the variables by sectors, as within a given sector, most of our variables share common sources. For each individual data variable, we provide information on its source, sample coverage, and, where applicable, the imputation procedures implemented.

3.1 Agricultural Sector

We have gathered 26 data series for the agricultural sector, with 20 of them reported annually and 6 reported at other frequencies.

Census data. We collected data on farm statistics from the NASS census,[6] spanning the period from 1870 to 2020. Data before 1920 are reported at ten-year intervals, while post-1920 data are reported in four- or five-year intervals. These farm statistics provided information on: [1] the value of crops produced; [2] the value of livestock on farms; [3] the value of forest products produced; [4] the value of farmland and buildings; [5] the expenditures on feed, fertilizers, and labor with board;

[6] The National Agricultural Statistics Service (NASS) is a component of the United States Department of Agriculture (USDA).

and [6] the number of farms (farm operations). We note that farm production expenditures on feed, fertilizers, and labor provide information on the value of intermediate inputs that were absorbed within the agricultural sector.

Crop, livestock and forest receipts. We constructed a measure of gross agricultural production by aggregating the cash receipts obtained from sales of crops, animals (including animal-related products), and forest products. Cash receipt data are reported annually by the ERS starting from 1924;[7] and these data are available in the "Value added by U.S. agriculture (includes net farm income)" table within the *Farm Income and Wealth Statistics* section. The ERS does not report state-level cash receipts for the period prior to 1924, offering only US-level estimates of annual cash receipts from 1910 to 1923. Henceforth, we refer to the annual cash receipts as United States Department of Agriculture receipts, or "USDA receipts".

[1], [2] and [3]—the Census-reported production of crops and livestock and forest products, respectively—are crucial in our imputation of annual cash receipts from 1870 to 1923. Our first task is to construct a consistent series for the value of crops, livestock and forest products at the Census frequency which can be spliced with the USDA receipts at the state-level from 1924 onwards. In particular, we are careful to address 1) the issue of double-counting in production value, which affects the early Census of Agriculture data and has been a primary area of focus for USDA economists from 1900 onwards, and 2) the difference between value of production and receipts (or value of products sold).

Double counting: The issue of double-counting is in essence that a portion of crops are purposed by farmers both as feed for animals, as well as seed for future planting. As such, a simple addition of the total production value of crops and livestock results in an overestimation of total farm production value. In comparison, the value of crops excluding double counting from feed and seed is termed in USDA reports as the "amount entering gross product".

This issue has been noted in the Census of 1870 to 1900, for which total value of production is reported alongside (and including) total value of livestock. In the Census of 1900 a significant effort was made for the first time to account for the double counting issue—Census surveyors collected at the state-level estimates of the value of crops fed to livestock, in addition to the total value of crops produced. However, as noted in Census 1910: "The sum of the values of [livestock and crop]

[7] The Economic Research Service (ERS) is another component of the USDA.

agricultural products involves a large amount of duplication, because great quantities of crops are fed to the animals on the farms. The amount of such duplication can not be measured accurately and the results of an attempt to do so in 1900 were not considered satisfactory. It has been thought best, therefore, not to attempt to give any estimate of the total value of agricultural products in 1909." We note that the individual crop production value reported in Census 1910, 1920 and 1925 do not adjust for double counting.[8]

In the years since 1909, great effort has been made by USDA statisticians and Census surveyors to fully address the double counting issue—at the state-level from 1924 onwards, and at the US-level from 1910 onwards, as well as backwards at the Decennial Census frequency back to 1800. From 1910 onwards, this is captured in the US-level USDA Value Added series. Pre-1910, Towne and Rasmussen (1960), as revised by Weiss (1993), estimate crop and livestock receipts adjusted for double counting at the US-level, and at the Decennial Census frequency.[9] We note that their work was precisely meant to construct estimates pre-1910 comparable with the post-1910 USDA US-level series.

Within our dataset, we build upon their work to construct state-level figures from 1870 to 1923 which adjust for double counting. In essence, we adjust the Census state-level crop and livestock figures by the US-level ratio of amount entering gross product, for each individual Census year. Specifically, for each Agricultural Census year from 1870 to 1925 we construct state-level shares using the Census reported production value figures, and apply these shares to the US-level receipts (sales) for that year, adjusted for double counting. We source the US-level receipts from Weiss (1993) for 1870 to 1900 decennially, and from USDA annually for 1910 to 1925. A welcome implication of this approach is that, for Census years, aggregating up our figures across states yields a national total which matches the Weiss (1993) figures.[10] Finally, we ratio-splice at the state-level with the state-level crop and livestock receipts from 1924 onwards.[11] In this manner, our data

[8] From 1870 to 1890, the value of crop production was not reported, only the total value of farm output, and the value of livestock. Thankfully, both the value of crop production as well as the total value of farm output is reported in Census 1900. We ratio-splice the two series, applying backwards the 1900 state-level ratio of value of crop production/total value of farm output (including that which is fed to livestock). Note that we do not exclude the amount fed to livestock here since 1) we are applying the share backwards to 1870-1890 for which the value fed to livestock is included 2) we account for the double counting issue in a manner detailed below.

[9] The revised aggregate figures were kindly shared in personal correspondence by Tom Weiss.

[10] After excluding Alaska, Hawaii and DC.

[11] In ratio-splicing the 1925 data, we note that the 1925 Census of Agriculture, being the first agricultural mid-decennial Census, reported a limited subset of crops and livestock compared to the decennial issues. That being said, all major crops and animals that we use in our Chow-Lin imputation were included. For livestock, the value reported in Census

attempts to address the double-counting issue.

Another concern that has been addressed with the above approach relates to the months of reporting. In particular, the crop data reported in the Census publications for crops relates to the crop year, while livestock data is reported at various dates. Crop years typically begin with the start of a particular crop's harvest, and end before the beginning of next year's harvest.[12] With the calendar year crop and livestock value recovered in Towne and Rasmussen (1960) from 1870-1900 and the USDA US-level receipts from 1910 onwards, our state-level estimates implicitly account for discrepancies in the months of reporting.

Commodity Receipts. We collect annual data on 10 crop commodities, 5 livestock commodities, and 1 forest product commodity.

Crop Commodities: We collect state and US-level data on 10 crop commodities from the NASS Quick Stats webpage: Oats Production Quantity and Prices, Wheat Production Quantity and Prices, Corn Production Quantity and Prices, Barley Production Quantity and Prices, Cotton Production Quantity and Prices, Cottonseed Production Quantity and Prices, Rye Quantity and Prices, Tobacco Production Value, Potato Production Value, and Sweet Potato Production Value. We note that Production Value is defined as the total quantity produced multiplied by the price per unit.[13] This is distinct from Production Receipts which is analogous to the crops sold, and excludes amount fed to livestock for example. Since crop commodity receipts are more informative of farmer's income and underlying economic conditions, we endeavor to estimate commodity receipts (in calendar years). Relying on newly digitized data from Strauss and Bean (1940), and Lucier (1986), we take a similar approach to addressing the double counting issues in the case of aggregate

1925 included the aggregate of the values of reported classes, as well as the estimated value of "unimportant classes ... namely, asses and burros, turkeys, ducks, geese, guinea fowls, pigeons and bees", this estimation being done based on the Census 1920 ratios. Thankfully, the Census also reported the total value in 1920 of the subset of crops reported at the mid-decennial Census. This allows us to follow the approach taken by Census for livestock to estimate the unreported crops—we construct ratios of the value of the subset of crops at 1920/total value of crops in 1920, and apply this ratio to the value of the subset of crops in 1925. A related concern in the calculation of gross product that has been addressed in later years by USDA statisticians is accounting for home consumption. They account for this at the state-level from 1949 onwards, and at the US-level from 1910 onwards. We do not adjust for this due to the length of the gap in the state data.

[12] Census 1925 relates to the crop years 1924. Livestock refers to the values on Jan 1 1925. Census 1910 relates to the crop year 1909. However, livestock refers to the count as of April 15, 1910. Census 1900 relates to the crop year 1899. However, livestock refers to the count as of June 1, 1900. Census 1890 relates to the crop year 1889.However, livestock refers to the count as of June 1, 1890. Census 1880 relates to the crop year 1879. However, livestock refers to the count beginning June 1, 1879, the enumeration going on until mid 1880. Census 1870 relates to the crop year 1869. However, livestock refers to the count beginning June 1, 1869, the enumeration going on until mid 1870.

[13] USDA reports quantities in crop years, values in crop years, and prices in marketing years.

crop receipts: For each commodity, we construct US-level ratios of receipts in calendar years to value in crop years, and apply these to the state-level values. In what follows, we describe in detail our treatment of individual crop commodities:

Oats, Wheat, Corn and Barley: We collect state-level (crop year) quantities, as well as state and US (marketing year) prices from NASS Quick Stats. We estimate state-level values by multiplying the state-level quantities by the state-level prices whenever available, and US-level prices otherwise.[14] We note that the reported prices are in marketing years, which is distinct from both crop and calendar years. A potential alternative to back out US crop year prices is to divide the crop year US values by the crop year US-quantities. However, we decide to use the marketing prices for two reasons: 1) we only have state-level prices for marketing prices, not crop year prices[15] 2) marketing prices only start differing from crop prices after 1900.[16] Next, for each commodity, we take US-level gross receipts in calendar years from Strauss and Bean (1940) from 1870-1910, ratio-spliced with Lucier (1986) from 1910-1960.[17] We construct an annual US-level ratio of gross receipts (in calendar years) to total value, and apply these to the state-level values to obtain an estimate of state-level receipts for each commodity.

Tobacco, Potatoes and Sweet Potatoes: We collect state and US-level (crop year) values from NASS Quick Stats. We note that there are select gaps where sweet potato values are not reported. For these periods, sweet potato quantities are reported, but not prices. In particular, sweet potato prices (both state and US level) are not reported for all states from 1876-1878 and 1881, 1892-1893, 1898. We back out the sweet potato value for these years by first imputing the prices for the missing years above (taking the average between the next and last reported years).[18] Finally, we multiply by state-level quantities to obtain value. We repeat the value to receipts procedure we use for Oats, Wheat, Corn and Barley, to obtain estimates of state-level commodity receipts.[19]

[14] We correct a likely error in the Quick Stats data for Maryland in 1890. The price reported is \$512/bu as opposed to \$0.38 in 1891 and \$0.3 in 1889. We set it to the average between these two years.

[15] Note that state-level value is not always available for all commodities, which means that we often cannot recover the crop year prices.

[16] In particular, marketing year prices tend to match crop years perfectly until around 1908, after which they start to diverge (for most crops by not more than \$0.05).

[17] Note that these are reported as "Gross Income".

[18] We choose this simple approach due to it generating less spikes compared to alternatives such as applying the $\hat{\beta}$ from a regression of sweet potato prices on other commodity prices.

[19] A slight difference is that, for sweet potatoes, our constructed US receipt to value shares from Strauss and Bean (1940) are spliced with those from Lucier (1986) in 1924 instead of 1910, as Lucier (1986) only starts to report them in 1924.

Cotton: We note that cotton refers to cotton lint (which makes up the majority share of value), as opposed to cottonseed. We take state-level quantities, as well as state and US-level prices from NASS Quick Stats. We note that, from 1869 to 1875, Quick Stats does not report Cotton prices (in marketing, crop, or calendar years). Thankfully, we are able to recover the calendar year prices digitized from Strauss and Bean (1940). In order to obtain calendar year quantities, we construct shares at the US-level using total cotton quantity in crop years (from Quick Stats), and total cotton quantity in calendar years (from Strauss and Bean (1940). We apply these to the state-level quantities in crop years and multiply by the calendar year prices to obtain estimates of state-level receipts from 1869-1875.[20] We perform this procedure for an additional year, 1876, in order to ratio-splice it with cotton receipts from 1876 to 1960, which are constructed in using the same value to receipts procedure we use for Oats, Wheat, Corn and Barley.[21] As a final note, Quick Stats reports Cotton Quantity as "480lb" bales throughout. However, this in fact changes across the years reported. From personal correspondence with USDA, changes that have occurred include: 1) Estimating running bales produced prior to 1899 (We note that running bales are the weight of a bale of cotton as it comes from the gin. This varies between 478lb and 506lb. For our purposes, we standardize it at 480lb.), 2) Estimating production of 500lb bales from 1900 to 1952, 3) Estimating 480lb bales from 1953 to current. When calculating cotton production value from quantity and prices (reported in \$/lb), we adjust accordingly.

Cottonseed and Rye: We collect state-level and US-level quantities from NASS Quick Stats. These crops are similar to cotton, in that Quick Stats does not report prices pre-1909. We perform the same procedure as Cotton to estimate state-level receipts: From 1875-1908, we construct shares using US-wide cottonseed quantity and the digitized Calendar year production from Strauss and Bean (1940). We impute state-level quantities from Quick Stats to Calendar years based on the ratio from Strauss and Bean (1940). We multiply by the digitized Calendar year prices for that period. We perform the same procedure for Rye from 1870-1908. For cottonseed, Strauss and Bean (1940) only report calendar year prices starting from 1875, and crop year data starting from 1874. From 1870-1874: We impute calendar year price in 1874 using the 1875 ratio and apply this price from 1870 to 1873, since the calendar year cottonseed prices are stable from 1875 to 1880. USDA

[20] We note that due to differences in units the calendar year prices in Strauss and Bean (1940) are 1/100 of the prices reported in Quick Stats (but otherwise identical for early years). We adjust accordingly.

[21] We apply the US-level ratios using calendar year receipts from Strauss and Bean (1940), ratio-spliced with Lucier (1986) from 1924-1960, since Lucier (1986) only starts reporting cotton receipts in 1924.

starts reporting US-wide cottonseed and rye prices (marketing years) from 1909. Therefore from 1909 onwards, we are able to impute the value from Quick Stats quantities and US-prices and follow the value to receipts procedure that we used for Oats, Wheat, Corn and Barley.[22]

Livestock Commodities: We collect state and US-level data on 5 livestock commodities, Cattle Value, Hogs Value, Sheep Value, Horses Value, and Mules Value, from newly digitized sources.

Cattle, Hogs and Sheep: We digitize and collect annual state-level production value (i.e. number multiplied by average value per head) from three sources: 1) 1869-1935: "Livestock on Farms, January 1, 1867-1935", 2) 1936-1939: "Agricultural Statistics" 3) 1940-1960: "Livestock and Poultry Inventory".[23] Since these individual series are all reported on Jan 1, and reflect the value in the prior year, we shift all years back by 1 before entry into our dataset. To account for slight revisions in the data in 1935, we ratio-splice the series when combining across sources. We note that: 1) We consider all types of cattle, including both milk cows and calves. 2) We only consider stock sheep (i.e. sheep reared for wool), excluding sheep and lambs on feed (i.e. sheep reared for meat) throughout. This is due to the fact that stock sheep made up the majority of the sheep in the early years, and were consequently much better reported on an annual basis. Finally, we estimate state-level receipts from state-level production values. We perform the same procedure as with crop commodities, employing newly digitized US ratios of receipts to production value: Strauss and Bean (1940) from 1870 to 1909, ratio-spliced with Lucier (1986) from 1910 to 1960. For robustness, we report an alternative measure of receipts for cattle and hogs, which has been ratio-spliced with the post-1924 USDA ERS data on cattle and hogs receipts whenever reported. There are two reasons that we do not use this measure as a baseline for all livestock commodities: 1) From correspondence with Monika Ghimire (USDA), we note that state-level data for most individual livestock commodities apart from cattle and hogs is only available from 2008 onwards. 2) As a secondary concern, there are gaps in the ERS data on cattle and hogs receipts for select states, while the value data is available in the original reports. In particular, this applies to data on cattle for New Hampshire (1955-1959) and Rhode Island (1950-1954), and data on hogs for New Hampshire (1955-1959), Vermont (1955-1959) and Rhode Island (pre-1945 and post 1948). In such cases,

[22] Note that due to differences in units the rye calendar year prices in Strauss and Bean (1940) are 1/100 of the prices reported in Quick Stats, but are otherwise similar for the early years. We adjust accordingly. We splice the ratios from Strauss and Bean (1940) with those from Lucier (1986) for rye in 1910, but cottonseed in 1924, as Lucier (1986) only starts to report them in 1924.

[23] Note that from 1945-1950 this is reported under the title "Livestock and Poultry Inventory on Farms and Ranches, Jan 1".

we fill in the gaps using our baseline measure of receipts which is constructed from the digitized value of cattle and hogs. Similarly, pre-2008 individual crop commodity data is only available for Cotton, Tobacco, Wheat and Corn, but not for Barley, Corn, Oats, Rye, Potatoes, Sweet Potatoes. Finally, we note that we perform our temporal disaggregation on both the baseline figures as well as the spliced figures, and the baseline results are robust to these alternative measures. In our main dataset, we report the spliced measures for Cattle and Hogs.

Horses and Mules: We note that, from Census data in 1870, horses and mules were the biggest category (representing 40% of livestock value for the US). We digitize and collect annual state-level production value from the same publications as Cattle, Hogs and Sheep. As we do not observe US-receipts for horses and mules, we report the production values. We note that from 1940 to 1960, only the aggregate value of horses and mules are reported. In order to impute individual horses and mules values from 1940 to 1960, we compute shares of individual horses and mules in 1939, and apply these to the data on aggregate horses and mules.[24]

Forest Product Commodities: We collect state and US-level data on lumber, for 34 individual species, decennially from 1869 to 1899 and annually from 1904-1945. This data is newly digitized from the publication: "Lumber Production in the United States, 1799-1946". While we only use aggregate lumber value for each state in our agricultural estimation procedure, we observe substantial heterogeneities in prices across both tree species and states. Since the publication does not report total value for each state, only total production, we digitize data on 34 species of trees in order to capture these heterogeneities in a more accurate aggregate value. In particular, 1) 1899, 1904-1945 (Annual): We collect data on state-level lumber quantities by state and tree species from 1904-1945, and state-level prices by state and type of tree for almost all years (prices not reported in 1905,1913-1914,1944-1946). For these select years where prices are not reported across all states, we impute prices as the average of the last and next reported values. Occasionally, there are states for which the state-level prices are reported only for some states that produce that type of lumber. In these cases, we take the US-price. 2) 1869-1889 (Decennial): For these years, prices are not reported. For 1889, we digitize state-level prices from the bulletins related to Census 1909. For 1869 and 1879, to the best of our knowledge, these prices are not available. However, "Lumber

[24] A notable omission from our 5 major livestock commodities is Poultry. We note that Poultry was not reported in census in 1870 and only starting in 1880. Further, it was not reported on an annual basis in the early years within the USDA documents. Therefore, it is our understanding that the 5 livestock commodities we report capture the lion's share of livestock value in the early years.

Production in the United States 1799-1946" gives us US-prices in 1859 and 1889, and we impute prices for 1869 and 1879 as the average US-price between these two years.

Next, we collected data on the number of farm operations from the NASS Quick Stats webpage, covering the period from 1910 to 2020.[25] According to the Quick Stats glossary, this number is intended to represent farms, farmers, ranchers, or growers within the agricultural sector, thereby providing an overview of the sectoral suppliers. We combine this series with the number of farms that we have collected from the NASS Census from 1870 to 1900. Additionally, we collect farm value per acre from the NASS Quick Stats webpage, which is decennial from period from 1870 to 1900, and annual from 1910 to 2020. Finally, we compiled estimates of farm real estate values, defined as the total value of the farm property class "land and buildings". Data from 1870 to 1949 was collected from Department of Agriculture publication "Farm Real Estate Historical Series Data, 1850-1970", and data from 1950 to 1995 from the USDA webpage, which primarily sources from the publication "Farm Real Estate Historical Series Data, 1950-92". In order to adjust the earlier years in line with the revisions made post 1950, we weight the pre-1949 years by the state-specific 5-year average from 1950 to 1955 of the shares of the data from "Farm Real Estate Historical Series Data, 1850-1970" and "Farm Real Estate Historical Series Data, 1950-92". Note that these shares are above 94%, and only affect growth rates for the year 1950. Finally, we collect data from 1996 to 2001 by multiplying the per-acre real estate value from the USDA publication "Land Values and Cash Rents" with the number of farms from "Farms and Land in Farms". We note that this method is consistent with how the total real estate values are calculated in the later publications of "Land Values and Cash Rents", from which we collect data from 2002 to 2008.[26] Finally, we collect data for 2009 to 2021 from the NASS Quick Stats webpage.

3.2 Mining Sector

We have gathered 6 data series from the mining sector, and providing information on: [1] the total value of mineral production (both metallic and nonmetallic minerals); [2] the quantity of petroleum production; [3] the quantity of coal production; [4] the quantity of gold production; [5] the quantity

[25] We note that this data is reported on 1 June from 1910 to 1992, and in calendar years from 1993 to 2020.

[26] We note that there was a revision to California's average real estate 2005 value of land and buildings from 4160$ per acre to 5090$ per acre in "Land Value and Cash Rents 2006". This, however, inflates the growth rate in 2005 by 0.19%. We note that the growth rates in the preceding and following 5 years are stable and consistently under 0.1. Given that this is the case, we smoothen the increase in the 2005 growth rate over the preceding and following 5 years.

of silver production; and [6] the quantity of pig iron production. The first variable is meant as a measure of the aggregate production value of the mining sector. The remaining variables were included because they were widely produced and reported in the early years. Specifically, coal and petroleum accounted for more than 60% of the total production value among the nonmetallic minerals from 1880–1930.[27] Similarly, the production of gold, silver, and iron accounted for most of the production value among the metallic minerals. We hope that these 5 main minerals may be helpful to paint a picture of the very early years, particularly before 1905 when there was a lack of comprehensive records for [1]. Moreover, we note that, as per Herfindahl (1966), in certain regions and states the mining industries have great influence on regional and state-level trends in economic activity.

Total Value of Mineral Production. From 1882-1904, we use a state shares approach. In particular, we use the US-level values of mineral production reported from the series Mineral Resources of the United States (Minres), a publication by the United States Geological Survey (USGS) that begins in 1882. We note that the USGS was established in 1879. We use this in conjunction with state-shares constructed using state-level data from Census in 1889 and 1902, as well as the 1905 Minres state-level data. In particular, our approach for the years 1890 to 1901 is to take the average of the shares between the Census years, and similarly between 1902 and 1905. We apply the 1889 shares to 1882 to 1888. From 1905-1919, data is taken from the state-level totals (i.e. aggregated over the minerals each state produces). However, one concern in the aggregation is a duplication of values, due to the fact that both raw materials and certain derivative materials are included. For instance, both the values of clay products and raw clay are included. Moreover, in order to prevent the disclosure of individual returns, Minres often does not report separately the values of some products under the respective states, and instead groups them under 'Other Products' or 'Miscellaneous'. This 'double counting' issue leads to a difference between the sum across individual state values and the US total (which does not include derivative materials), where the 'naive' total obtained by summing up the state totals is greater than the reported US totals which does not suffer from 'double counting'. We therefore construct state-shares for each year and apply them to the reported US total to obtain our 1905-1919 estimates.

From 1920 onwards, data are reported by the Census Bureau of Mines, as published in the SA.

[27] The share of contribution was approximately 74% in 1885, 65% in 1895, 61% in 1905, 62% in 1915, and 62% in 1925.

Notably, after 1988 reports on the total value of mineral production were split among two agencies – USGS reports only non-fuel mineral production as published in the Mineral Yearbooks, and EIA reports data on fuels. Accordingly, SA stops reporting total value of non-fuel + fuel production. They instead only report non-fuel mineral production, from USGS, and from EIA, data on value for the major fuels: Petroleum, Natural Gas, and limited data on Coal. Our approach is therefore to 1) Obtain total value of non-fuel mineral production from SA (1988-2001), USGS Statistical Summaries (2002-2019), and USGS Mineral Commodities Summaries (2020-2021). 2) Obtain total value of fuels by combining the total value of Petroleum production + total value of Natural Gas production + total value of Coal production. 3) Sum up the value of fuel and nonfuel mineral production to obtain a series comparable with the pre-1987 data.

[1] We obtain total value of petroleum combining the EIA series: State-specific Field Production of Crude Oil (Thousand Barrels) $*$ State-specific First Purchase Price (Dollars per Barrel). This series corresponds to the Value of Crude Petroleum as reported in SA.[28] [2] We next obtain the total value of natural gas by combining the EIA series: Natural gas production per cubic ft. $*$ Natural gas well-price per cubic ft. In a similar manner to Petroleum, this value series matches what is reported in SA for the overlapping years. One caveat is that EIA stops reporting well-price in 2010 (SA stops reporting after 2010 as well). As a proxy for well-price, we begin with Industrial price (as opposed to consumer price) as it matches well-price most closely for the overlapping period (1997-2010). Then, we construct the shares of (well-price/Industrial price) for 1997-2010. Since these shares are stable, we then compute the state-specific 5 years average share from 2006-2010, and extend the "well-price" series from 2011-2021 by taking Industrial price$*$share. We then construct Natural Gas value series by multiplying with the Natural Gas production series. [3] We obtain the total value of Coal by combining the EIA series: Coal price per BTU $*$ Coal Production in BTU. Finally, we sum up these three series along with the total value of production for non-fuel minerals (from USGS), in order to obtain the total value of Mineral Production.

Quantity of Petroleum produced. Data from 1876-1975 are taken from Bureau of Mines Mineral Yearbooks (MinYears). Data in 1976 are taken from Energy Information Administration

[28] Note that for a small number of states that produce relatively little petroleum, we have data on petroleum produced but not a state-specific first purchase price. In such cases, we use the US price for that year.

(1976). Data from 1977 - 1979 are taken from the SA and MinYears.[29] Data from 1981 - 2020 are taken from Energy Information Administration: Open Data. In particular, the state-level series Field Production of Crude Oil (Thousand Barrels).

Quantity of Coal produced. From 1871 to 1879, and 1881-1882 we use values from the USGS COALPROD database, which for this period sources data from an unpublished manuscript by Christiansen (1948) which had been obtained from EIA. This data encompasses 21 major Coal producing states. For 1870 and 1880, we supplement the data for the states not covered using the Decennial Census. From 1877 to 1879, we obtain the Michigan Coal statistics from the "Annual Report of the Commissioner of Mineral Statistics of the State of Michigan", a publication which begins in 1877. For Iowa, Kansas, Missouri and Washington from 1871 to 1879, and Michigan from 1871, we impute the data by first computing the shares in 1870 and 1880 of the coal quantity in the given state/the total quantity of 14 states consistently reported from 1870 to 1880. We compute this share in 1870 and 1880, and take the average of these shares. Finally, we apply this share to the total quantity of 14 states consistently reported, for each year between 1871 and 1879. From 1883-1976, we obtain data on Coal Production as reported in the MinYears, and as published in the SA.[30]. From 1977-1979, we again use values from the COALPROD database, which corresponds with the EIA data. Finally, from 1980-2020, data is compiled from the EIA Annual Coal Reports.

Quantity of Gold produced. Our primary source for Gold production comes from Craig and Rimstidt (1998), which reports state-level data from 1800-1995. The paper's main sources are in turn, in order of usage, the Bureau of Mines Annual Reports, various USGS reports, as well as some state-specific reports. From 1996-2020, we obtain state-level data from MinYears 1996-2020.

Quantity of Silver produced. From 1877-1927, we collect data from the report Merrill (1930), which is a joint publication between the Bureau of Mines and the Department of Commerce. From 1928 until 1943, we report the quantity as published in the Statistical Abstract of the United States (SA). The primary source of these figures is from the Annual Reports of the Director of

[29] The Bureau of Mines was dissolved in 1996, and succeeded by the USGS. After this change, the USGS continued to publish the Mineral Yearbooks from 1996-present.

[30] Note that for the 21 major Coal producing states, the data corresponds with what is reported in COALPROD. When states produce a small amount of Coal, they are sometimes not reported in COALPROD but are in MinYears, and in such cases, we use the MinYears data.

the Mint, a branch of the Treasury Department. From 1944-2020, we report data from MinYears. In harmonizing the data, we note that a fine ounce is equivalent to a troy ounce as a measure of weight, with the additional requirement that the material in question must be sufficiently pure.

Quantity of Pig Iron produced. Data from 1878-1967 is reported from the Annual Reports of the American Iron and Steel Institute and the Bureau of Mines, as published in the SA. From 1968-2020, data is compiled from MinYears.

3.3 Manufacturing Sector

We have gathered 4 data series from the manufacturing sector, providing information on: [1] the total number of manufacturing establishments; [2] the total number of manufacturing employees; [3] the total value of wages and salaries paid to manufacturing employees; and [4] the total value added from the manufacturing sector. The first variable tracks the sectoral size, whereas the second and third variables provide indications of the sectoral labor force. Variables [2] and [3] are useful to gauge average income per worker, thereby facilitating the observation of sectoral income growth over time. Finally, the variable on sectoral value added provides insights into the sector's overall contribution to the state-level economy.[31]

[31] There are two definition changes of note in the early years, which affect all four of our variables. First, the 1900 statistics were revised in 1905 for purposes of comparability, however, the statistics for 1870, 1880 and 1890 were not revised. This affects all four manufacturing variables, but the number of manufacturing establishments in particular. The unadjusted 1900 statistics were reported in the 1900 census, and the adjusted 1900 statistics in the 1910 census. In order to preserve the comparability of the 1870, 1880 and 1890 statistics, we constructed a state-level ratio between the adjusted and unadjusted data in 1900 and applied them to the 1870, 1880 and 1890 unadjusted statistics. A second definitional change of note is, in 1921, a change in the establishments from which data was collected. In contrast to earlier years, from 1921 onwards, data was not collected from establishments reporting products valued between $500 and $5000. Accordingly, we see a dip in the raw data as reported over all four variables. Thankfully, data was reported broken down over value of products reported by establishments was detailed in the census of 1905, 1909, 1914, and 1919. For example, it reports the number of employed in establishments with products valued "Less than $5000", "$5000 and less than $20000", "$100000 and less than $1000000", and "$1000000 and over". In particular, such breakdowns are available for number of establishments, employed, and value of products, but not total payroll. We therefore report the respective variables excluding the "Less than $5000" bin, to cohere with the post-1921 definitions. Payroll was not reported at this level of detail, however, given that payroll and number of employed are highly correlated with each other, we apply the share of employed in total establishment worth less than $5000 over the total employed in all establishments to our data on wages. For 1870, 1880, 1890, and 1900, in lieu of value of product level breakdowns, we apply the average of the 1904 and 1909 shares for each variable, to the preceding years. Additionally, we note that the automobile repair industry was not surveyed in the 1921 ASM, but was reported in previous years. We do not adjust for this for two reasons: 1) it accounts for a small percentage of the total manufacturing industry—0.4% of value added in manufacturing in 1919, 2) the industry-state-establishment bins by value of products is not consistently reported.

Establishment counts. The number of establishments was culled from the *Census of Manufactures* (CM) and the *Annual Survey of Manufactures* (ASM).[32] More specifically, we collected establishment counts from the *Census of Manufactures* at ten-year intervals from 1870 to 1900, at five- or four-year intervals from 1905 to 1915 and from 1954 to 2017, at two-year intervals from 1919 to 1939, and for a single year in 1947. From 1919 to 1939, we addressed missing data by averaging establishment counts over adjacent years. For instance, we imputed the establishment counts in 1920 by averaging the numbers reported for 1919 and 1921. In so doing, we made the assumption that establishment counts did not experience significant fluctuations between consecutive years over the two decades. Next, we describe the steps taken to impute annual establishment counts from 1940 to 1951. For the years spanning 1946 to 1951, we first derived state-specific year-on-year growth rates for establishment counts based on data from the *County Business Patterns*.[33] Then, we applied the growth rates to estimate annual establishment counts for the years 1946 and 1948–1951 based on the 1947 data obtained from CM. Finally, for the WWII period (1940 to 1945), we imputed annual counts by linear interpolation using data for 1939 and 1946. While the interpolation-based method is relatively simple, it yields estimates that align with the broader observations of increased manufacturing activities in the U.S. during the WWII period; for instance, see Jaworski (2017). Starting from 1953, we collected intercensal establishment counts from the ASM to supplement the CM data. We employed CBP growth rates to impute these post-1953 data whenever ASM data were missing. We note that there were four years (i.e., 1952, 1957, 1960, and 1961) for which no observations were reported in the ASM, CM, and CBP. In these cases, we imputed the missing data either by employing linear interpolation across years or by calculating averages using observed data.

[32] Both CM and ASM were published under the Bureau of the Census. CM data are reported in ten-year intervals from 1870 to 1900 (i.e., for years 1870, 1880, 1890, and 1900), in five- or four-year intervals from 1905 to 1914 (i.e., for years 1905, 1909, and 1914), and in two-year intervals from 1919 to 1939 (i.e., for years 1919, 1921, ..., 1937, 1939). The CM was discontinued during the WWII period. It resumed for a year in 1947 before transitioning to reporting in five-year intervals from 1954 to 1992. Subsequently, the CM is reported as part of the *Economic Census* after 1992, from which we collected data for the years 1997, 2002, 2007, 2012, and 2017. The first issue of ASM was published in 1949, complementing the CM by providing intercensal data annually. Accordingly, the ASM and CM data are directly comparable; see Census (1963) for a detailed comparison. Hence, we do not face definitional issues when constructing a variable using both the ASM and CM concurrently.

[33] The *County Business Patterns* (CBP) series reports county-level manufacturing data annually starting from 1946 (with certain years missing). We aggregated county-level data to the state level and observed discrepancies when compared to the state-level data reported in the CM and ASM. These discrepancies may arise from variations in sampling methods. Thus, we opted not to use the CBP values directly; instead, we employed them to infer year-on-year growth rates.

Employee counts. The number of manufacturing employees includes all full-time and part-time persons on the payrolls of reporting establishments who received compensation. This data series was not reported in the CM prior to 1939; and thus, we constructed the pre-1939 counts by summing up the number of "wage earners" and "salaried workers" in the census years. We note that this aggregation method is consistent with the census convention for the post-1939 data. Pre-1949, we first collected data from the CM, and this data therefore being restricted to the reporting frequencies as the "establishment count" variable. We proceed to extend the employment data using the employment index in Wallis (1989), which reports a state-level manufacturing employment index from 1929 to 1940. In particular, the data is reported as an index such that 1929 = 100. Since the data is reported as an index, we first construct year-on-year growth rates. We apply these growth rates backwards on the years 1931, 1933, 1935, 1937, 1939, to obtain manufacturing employment levels in 1930, 1932, 1934, 1936 and 1938. We apply the 1940 growth rate forward to the 1939 CM data to obtain manufacturing employment levels in 1940. From 1941–1946, we collect data from the BLS publication "Employment and Payroll". Whenever possible, we report data collected in December of the year (this applies to the years 1942–1945). In 1946, we were only able to obtain data in November. In order to account for potential seasonal changes from month to month, we weight the November 1946 values by the state-specific ratio of the December 1945/November 1945 values. Similarly, we proxy the employment in December 1941 by using the available January 1942 values, and weighting them by the state-specific ratio of the December 1942/January 1943 values. Post-1946 data were collected from the ASM annually, with four exceptions over episodes 1949–1951, 1952–1953, 1979–1981, and 1996–1999. In the first case (1949–1951), employee counts for the South and North Dakota states were reported as a whole in the ASM. We disaggregated these numbers using a two-step approach: First, we computed the employee share of North Dakota in year t, $s_{N,t} = \frac{Emp_{N,t}}{Emp_{N,t}+Emp_{S,t}}$, for $t \in \{1947, 1952\}$,[34] which were shown to be relatively stable across time (i.e., $s_{N,1947} = 34\%$ and $s_{N,1952} = 34\%$). Then, we backed out the employee count of Dakota states in year $t' \in \{1949, 1950, 1951\}$ by $Emp_{N,t'} = (\frac{1}{2}\sum_t s_{N,t}) \times Emp_{t'}$ and $Emp_{S,t'} = Emp_{t'} - Emp_{N,t'}$, respectively. In the second case (1952–1953), estimates of employee counts were not reported for Florida, and were therefore imputed using the total employee counts for the South Atlantic region. Finally, employee counts data were not reported in the ASM for the last two cases (1979–1981 and

[34] The 1947 data were taken from the CM, while data for 1952 were taken from the ASM.

1996–1999). Thus, in these periods, we employed the CBP growth rates to recover the missing data.

Wages and salaries. Data sources and imputation methods for wages and salaries are the same as the "employee counts" variable, with two exceptions: 1) in 1946 when no wage and salary data were reported in ASM, CM and CBP 2) when data is reported biannually from 1929 to 1939, we linearly impute the intervening years as wage and salary data is not available in Wallis (1989). We note that wages and wage earners account for on average 80% of the total manufacturing workforce and payroll, with salaried workers and salaries making up the other 20%. Additionally, while wage and wage earners data is reported in 1925, 1927 and 1931, salaried workers and salaries are not in 1931 and only reported on an individual industry-state level in 1925 and 1927. We therefore perform linear imputation for salaries and add them with wages for these three years. In this case, we imputed the missing data by linear interpolation based on data for 1939 and 1947. For 1981, we impute payroll data using payroll/employment shares constructed from the 1980 and 1982 data, and combine this with the 1981 employment figures reported in ASM 1981.

Value added. The total value added from the manufacturing sector are defined as the gross sectoral output less cost of production. For the period 1870–1978, we note that the sources and imputation methods for value added are the same as the "wages and salaries" variable; and thus, we omit the discussion in this section. Value added figures were not reported in the ASM, CM and CBP for 1979, thus we imputed by linear interpolation using data from adjacent years. For 1981, data on value of shipments is available but not total value added. We therefore construct shares of value added/value of products shipped for 1980 and 1982, and apply these shares to the 1981 data on value of shipments to impute value added for 1981.

3.4 Transportation Sector

We have gathered 5 data series from the transportation sector, reported annually from 1942 to 2020. These variables provide information on [1] the mileage owned of railroad tracks; [2] the total of rural and municipal road mileage; [3] the total state highway mileage; [4] the total number of motor vehicle registrations; [5] the total automobile-related revenues collected. Broadly, we report data on the transportation sector in an effort to capture the degree of connectivity within each

state, as a potentially useful economic indicator.

Mileage owned of railroad tracks. Data from 1885 to 1973 was compiled from SA.[35] From 1885-1904, data was originally reported in the "Poor's Railroad Manual". From 1870-1884, we digitize the data directly from "Poor's Railroad Manual". From 1905-1973, it is as originally reported in the annual statistical reports of the Interstate Commerce Commission (henceforth abbreviated as ICC).[36] From 1870 to 1888, the data for what became North and South Dakota was reported jointly under "Dakota Territory". Similarly, from 1870 to 1884, Maryland and District of Columbia were reported together. Since the post-1889 and post-1885 shares respectively remain stable, varying by less than 1% in the proceeding three years, we calculate the ratios of railroad mileage owned in 1885 for $\frac{Rail_{MD}}{Rail_{MD}+Rail_{DC}}$, and 1889 for $\frac{Rail_{ND}}{Rail_{ND}+Rail_{SD}}$ and $\frac{Rail_{SD}}{Rail_{ND}+Rail_{SD}}$, and then apply this ratio to the annual totals in the preceding years to back out the state-level data. From 1974 onwards, while data on mileage owned of railroad tracks continues to be reported in the annual statistical reports of the ICC, it is unfortunately reported on a railroad company-level instead of the state-level. Since railroad companies can operate across states, it is challenging to recover the state-level values.

Total rural and municipal road mileage. Data from 1904-2020 is collected from SA. From 1904-1930, data was originally reported within the Annual Reports of the Bureau of Public Roads, part of the Department of Agriculture. However, from 1931-1941, data on total rural and municipal road mileage is not reported within the Annual Reports of the Bureau of Public Roads (henceforth abbreviated as BPR), and neither is it reported within the SA. From 1942 onwards, data continues to be reported on an annual basis in SA, as originally sourced from the annual statements on the status of highways of the Public Roads Administration of the Federal Works Administration and the Annual Reports of BPR, and is reported as such.[37] We note that only data on rural mileage, not data on rural and municipal mileage is available before 1952. We ratio-splice this series with

[35] We note that we code two observations as potential reporting errors in SA due to unreasonably large single-year spikes considering the pre and post trends: Massachusetts in 1887 (reported 2025.87 in 1886, 3067.6 in 1887 and 2074.32 in 1888) and Iowa in 1895 (reported 8508 in 1894, 5523 in 1895 and 8511 in 1896.

[36] We note that 1942 data is not reported for all states, and we linearly impute accordingly.

[37] Note that the BPR was transferred in 1939 from the Department of Agriculture to become the Public Roads Administration of the Federal Works Administration. In 1949 the Federal Works Administration was abolished and the Public Roads Administration was renamed the BPR and transferred into the Department of Commerce. Therefore the two reports mentioned above, and the data within as reported within SA, can be seen as succeeding publications.

the post-1953 data on rural and municipal mileage post-1953.

State Highway Mileage. Annual data from 1923-2020 is taken from SA. Of note, 1935,1958 and 1973 are not reported in SA, and we linearly impute for these years. From 1923-1932, existing State highway mileage is reported as the total road mileage in state highway systems. 1933-1941 and 1953-1991, existing State highway mileage is reported as the existing mileage of rural roads and urban extensions of State Highways. Existing State highway mileage from 1945-1952 is not reported in SA, and instead compiled directly from "Highway Statistics" (henceforth abbreviated as Highstat), a publication by the Federal Roads Administration, a branch of the Federal Works Agency. Finally, data from 1992-2020 is reported in SA as "Other Arterial" highway mileage.

Number of Registered Motor Vehicles. Annual data from 1900-1995 is collected from Table MV-201 provided by the U.S. Department of Transportation. Data from 1996-2021 is collected from Highstat. Given that automobiles have remained significant as a consumer durable since 1913 up till the present, we report the number of registered motor vehicles, as well as the total automobile-related revenues in the hope that it may capture broader trends in consumer durables.

Automobile-related revenues collected. Data from 1913-1955 is collected from SA, as reported by the Bureau of Public Roads, Department of Agriculture. Data from 1956-2020 is collected from Highstat. Of note, 1921 and 1976 are not reported in SA or Highstat, and we linearly impute for these years.

3.5 Business Statistics

In this section, we focus exclusively on data concerning commercial failures and bankruptcies. There are two reasons that inform this choice. First, data on commercial failures have been consistently reported since the late nineteenth century; and second, commercial failures, along with the liabilities of failed firms, have been recognized as leading indicators of economic crisis; see Simpson and Anderson (1957) and Richardson and Gou (2011*a*) for details.[38] Our data set makes a distinction between commercial failures and bankruptcies. As noted in Simpson and Anderson (1957), a commercial failure is defined as a business concern that is involved in a court proceeding voluntarily

[38] For instance, Table 1 in Simpson and Anderson (1957) shows liabilities of failed firms led changes in business conditions (measured by peaks and troughs of industrial production) in nine out of twelve instances from 1929 to 1949.

or involuntarily, and which is likely to end in losses to creditors. We note that data on commercial failures were sourced from Dun & Bradstreet (henceforth D&B) and its predecessors.[39] However, D&B does not include data on failures of small businesses and personal bankruptcies. Therefore, we collected another series of bankruptcy data – as compiled by Hansen, Davis and Fasules (2016) – that encompasses both firm-level (of any size and types) and personal-level bankruptcies. Finally, we gathered data on the total number of business concerns, which allow us to compute failure and bankruptcy rates over time.[40]

D&B data. We have gathered 3 data series from the D&B publications, namely: [1] the total number of business failures; [2] the total business liabilities of failed firms, which refer to all current liabilities except for long-term publicly-held obligations such as bonds (this exclusion is comparatively small, averaging about 1% of total liabilities); and [3] the total number of business concerns. These data series are culled annually since 1885. For the first two data series, we have collected them directly from D&B publications whenever possible. In cases where copies of D&B publications were not easily accessible, we resorted to collecting them from the Statistical Abstract of the United States (henceforth abbreviated as SA).[41] Note that Dun & Bradstreet defines business failures as businesses involved in court proceedings or voluntary action involving losses to creditors. This excludes business discontinuances—for example, a business that discontinues operations for reasons such as loss of capital, inadequate profits, ill health or retirement. We note that data on business failures and liabilities of failed firms were collected from SA for the period 1885–1893. These data were originally published in the Annual Circulars of R. G. Dun & Co., and were not accessible by our team during the data collection process. Then, with the exception of 1933, we sourced business

[39] Dun & Bradstreet was founded after a 1933 merger between R. G. Dun & Co. and J. M. Bradstreet & Son, both of which were reporting agencies that collected data on business failures. We refer readers to Richardson and Gou (2011*a*) for a comprehensive account of the various venues and forms of Dun & Bradstreet's publications.

[40] We note in passing that differences in failure and bankruptcy rates at any time may not be reflective of the changes in *current* business cycle. This can be due to several reasons. First, there could be substantial time lag between firm default and bankruptcy filling (Hansen and Ziebarth, 2017), and thus, bankruptcy rates alone may be insufficient to capture the general positions of current business cycles. Furthermore, regulations such as the garnishment law could influence the degree of counter-cyclicality of bankruptcy rates. That is, bankruptcy rates may exhibit strong countercyclical patterns in states with stricter pro-creditor garnishment laws only (Hansen and Hansen, 2012). Nonetheless, we believe that failures and bankruptcies remain useful in our analysis. This is because, as aforementioned, economy-wide increase of failures and bankruptcies could play an important structural role in the propagation of *future* recessions.

[41] SA began reporting data on commercial failures by states in 1889 (see Table 8 of the 1890 SA). However, SA reports often rounded up dollar amounts of liabilities. For example, liabilities were rounded up to $\$1,000$ from the 1922 SA and to $\$1,000,000$ from the 1973 SA. To ensure completeness, we aimed to source our data from the original D&B publications whenever possible.

failures and liabilities data directly from *Dun's Review* for the period 1894–1932, from *Dun and Bradstreet's Monthly Review* for the year 1934, and later from *Dun's Statistical Review* for the period 1935–1957;[42] It is worth noting that our datasets align with those reported in Appendix 1 of Richardson and Gou (2011*a*) for the period 1896–1936.[43] Lastly, business failures and liabilities data for the year 1933, as well as post-1957 ones, were collected directly from the SA.

While annual data on business failures and failed liabilities have been consistently reported since 1885 for most states, we encountered three exceptions. First, Oklahoma did not commence reporting until 1889. To compensate for the gap in 1885–1888, we used data from Indian Territory as proxies for Oklahoma prior to 1889. In fact, we included Indian Territory data into Oklahoma's series until 1908 when Indian Territory ceased to be reported, this date corresponding with its entry as part of Oklahoma into the union. Second, Alaska and Hawaii did not start reporting annually until 1975. Third, data were not divided between the two Dakota states in the period from 1885 to 1887. In this regard, we apportioned the total figure using state shares inferred from the 1890 data.[44]

Next, we sourced the total number of business concerns from SA for the period spanning 1885 to 1968. D&B define the total number of business concerns as the total number of business enterprises listed in the D&B Reference Book.[45] In particular, each branch of a company is represented under a separate entry, and branches are listed geographically at the state-county-town level. The business enterprises listed within the book include manufacturers, wholesalers, retailers, building contractors

[42] Some of these D&B publications were published on a monthly basis. In this regard, D&B generally reported a table containing the details for the previous year in the January or February issues. We collected these annual data reported therein whenever applicable. Note also that Dun's Statistical Review only reports liabilities in thousands, and therefore our compiled data does as well.

[43] Richardson and Gou (2011*a*) also constructed a series that tracks the total number of business failures at the US-level from 1895 to 1940. For sanity check, we aggregated our state-level estimates to national figures and compared against those reported in Richardson and Gou (2011*a*). Our US-level series aligned well with those reported by the authors, and in particular, from 1932 to 1940, our figures differed by at most 4% from the numbers presented in their paper. The one exception to this was in 1936, where the authors reported a large spike of 26255 failed businesses compared to 1935 (11879 failed businesses) and 1937 (9217 failed businesses), while our figures did not, i.e., we reported 11510, 9185, and 9490 failed businesses for 1935, 1936, and 1937, respectively. We note in passing that these original D&B publications were sourced from a variety of sources, including the HathiTrust webpage (1894–1922), the off-site facility within the New York Public Library (1923–1931, 1935–1957), and the closed stacks within the Peking University library (1932, 1934).

[44] We computed state shares using the 1890 data because no data were reported for the Dakota states in 1888 and 1889. Furthermore, the year-on-year state shares exhibited substantial fluctuations from 1890 onwards, making it impractical to rely on averages in our imputation process. By fixing the state shares of 1885–1887 to the level of 1890, we are at best offering rough insights of the business failures and failed liabilities in the two regions before 1890.

[45] We note that the D&B Reference Book has been published as early as 1859, detailing a listing of business concerns and their creditworthiness. However, the Reference Book does not provide a count of the businesses within at the state-level, only at the national level at the beginning of each book.

and certain types of commercial services including public utilities, water carriers, motor carriers and airlines.[46] A similar imputation process was applied to recover the Oklahoma figures for the years 1885 to 1888, and to split the Dakota figures into North and South Dakota for the years 1885 to 1887.

One definition concern of note has to do with the passing of section 77B in 1934, which allowed for a new form of corporate reorganization. As detailed in Richardson and Gou (2011*b*), section 77B altered the nature of bankruptcy, and statistics on bankrupt businesses. In particular, Dun and Bradstreet report data on the number of 77b applications from Jul 1935 to Sep 1938. In 1938, the Bankruptcy Act of 1938 (also known as the Chandler Act) eliminated section 77B, and corporate reorganizations now occurred under Chapter X of the Chandler Act.[47] The cases filed under chapter X were subsumed within the Dun's failure series.[48] In sum, we have three series that are related to this definition change: 1885-1934 without section 77B or Chapter X, 1935-1937 with section 77B, and post-1937. We note that Richardson and Gou (2011*b*) construct a US-industry-level series from 1895 to 1939 excluding section 77B to be consistent with the earlier year data. Since we are focused on longer time horizon, do not distinguish between industries, and are primarily concerned with the year-on-year growth rates, we ratio-splice the series (including 77B filings) with the pre-1934 series (before the formation of 77B). In particular, we construct a US-wide ratio in 1934 of failures without 77B applications/failures with 77B applications, and ratio splice based on this.[49]

We note that in January 1935, Dun and Bradstreet's failure statistics were revised to exclude failures of insurance and real estate agents and brokers, holding and finance companies, shipping agents, tourist companies, transportation terminals and such. These revisions brought the failure record in alignment with the number of concerns reported, for which no changes were made. We report the figures as revised from 1935 onwards. We ratio-splice with the pre-1935 series at the

[46] However, this count does not cover all the business enterprises of the country. Specific types of business not listed are: financial enterprises including banks, mortgage, loan and investment companies; insurance and real estate companies; railroads; terminals; amusements; and many small one-man services. Note that it also excludes many small one-man services, consistent with how the listed total number of business failures excludes small businesses.

[47] For more details on the Chandler Act, see the August 1938 issue of Dun's Review.

[48] See the Feb and March 1939 issues of the Dun's Review.

[49] The data on 77B applications in 1934 is taken from Dun's Review April 1937, pg 36. An alternative approach to imputation is to construct state-level shares using state-level 77B applications in 1937 and 1938 from Hansen, Davis and Fasules (2016), combined with US-level 77B applications for the Dun and Bradstreet sample from Richardson and Gou (2011*b*) table 14, and US-level failures excluding 77B from Richardson and Gou (2011*b*) table 6. We do not take this approach as there is often great variation in state-level growth rates of failures from year to year to the small levels values, and the 1937 shares backwards may generate noisy levels values. This is for liabilities in particular, since we do not have data on liabilities of 77B applicants, we would like to impute using the number of failures.

state-level, taking advantage of the fact that adjusted and unadjusted state-level figures for 1935 were reported in SA 1936. As a robustness check, we compare the spliced figures for 1933 and 1934 to the reported adjusted US totals (state-level adjusted totals were not reported). These differ by less than 1% for both years. We choose to ratio-splice using the 1935 state-level ratios as opposed to applying the 1933 or 1934 US-wide ratios backwards. The reason for this is that, comparing the 1935 adjusted and unadjusted values, there is substantial heterogeneity across states – in particular, 30 states are unaffected by the adjustment, while the affected states experience up to a 6% difference.

Similarly, beginning 1939 Dun and Bradstreet extended their sample to include voluntary discontinuances with loss to creditors and small concerns forced out of business with insufficient assets to cover all claims. We ratio-splice in a similar manner to the 1936 definition change, using the 1939 adjusted, and unadjusted figures as reported in SA 1940.

Finally, Dun and Bradstreet extended the coverage of failure series to more sectors from 1984 onwards. In particular, coverage was extended to include the sectors agriculture, forestry and fishing, finance, insurance, and real estate, all miscellaneous services beyond repair services, public administration, and businesses for which their industries were deemed not classifiable. Note that the latter three categories are reported under 'unclassifiable' in the Dun and Bradstreet reports. As demonstrated in Naples and Arifaj (1997) for the US as a whole, naively joining the series before and after 1984 can lead to misleading results. They address this discontinuity and create a consistent US-level failure series by dropping the newly added industries from 1984-1998.[50] This takes advantage of the fact that Dun and Bradstreet report the failures data at the industry-level. We extend their work by splicing the series at the state-level, using newly digitized state-industry data as published by Dun and Bradstreet.

Post 1998, we extend the series using Business Filings under the Bankruptcy Code, and ratio-splicing the two series. As with our other bankruptcy variables, data from $1998 - 2007$ was taken from Hansen, Davis and Fasules (2016), while data from $2008 - 2021$ was taken from Table F-2 of the December U.S. Bankruptcy Courts reports.

[50] Note that we use the figures as reported from 1984-1996. From 1997-1998, we apply the preliminary 1997 ratio extracted from the 1997 report to the total failures as reported in SA.

Bankruptcy data. Note that we consider the universe of bankruptcies which includes both personal and corporate bankruptcies. From $1900-2007$, we obtained the total number of bankruptcies commenced, and the total number of bankruptcies closed for each fiscal year from Hansen, Davis and Fasules (2016). There have been multiple changes in Bankruptcy law, most prominently the change from the Bankruptcy Law (pre 1980) to the Bankruptcy Act (post 1980). These changes correspond to different variables as reported in Hansen, Davis and Fasules (2016). Bearing in mind that we are primarily concerned with year-on-year growth rates instead of levels, we attempt to form a consistent time series across our sample by ratio-splicing.[51] From $2008-2021$, we collected these two variables from Table F of the December U.S. Bankruptcy Courts reports. We join the data with the pre-2008 data, ratio-splicing using the base year 2007.

3.6 Government Finances

We have gathered 8 data series that deal with government finances, which are reported annually from 1870 to 2020. These variables provide information on government revenue, government expenditure and government debt. Namely: [1] the total state government general revenue collected; [2] the total federal government internal revenue collected; [3] the total personal income tax; [4] the total corporate income tax; [5] the total state government general expenditure; [6] the total long-term state government debt; [7] the total gross state government debt; [8] the total net state government debt. Together, revenue, expenditure and debt provide a complete picture of state government finances.

Total Federal government internal revenue collected. Internal revenue is a part of Federal government revenue. Accordingly, our state-level measure of internal revenue collected reflects the amount of internal revenue collected, on the behalf of the Federal Government, within a particular state. The major components of internal revenue are individual and corporation income taxes, social security taxes, and miscellaneous internal revenue taxes. Notably, before the introduction of income tax in the form of the Sixteenth Amendment in 1913 (and collected from 1915 onwards),

[51] Since we do not have overlapping years, we apply a 'reasonable' growth rate backwards, constructing this growth rate as the mean of the 5 years following the definition change. Note also that the variables in Hansen, Davis and Fasules (2016) are often only available under subcategories for certain years. For example, the years voluntary and involuntary cases commenced are available is a superset of the years total cases commenced is available. Since these subcategories sum up to the total for the years both are available, we often reconstruct the totals in the years only subcategories are available.

internal revenue was primarily made up of what later became known as "miscellaneous internal-revenue taxes". This included sales taxes on alcohol, oleomargarine, stamp taxes, etc.

After the introduction of income taxes, individual and corporation income taxes became the largest component of internal revenue. Given that this is the case, we also report them separately. There are two reasons that we decide to report the breakdown for income taxes for federal government internal revenue, but not state government general revenue. First, federal income taxes (i.e. collected for the Federal Government) were imposed on all states starting from 1915; in contrast, the adoption of state income taxes (i.e. collected for individual state governments) was slower. In 1915 only 5 states reported revenues from an income tax, 9 states in 1919, 13 states in 1925, 28 states by 1933 and 34 states by 1938. Therefore federal income taxes allow for better comparability across states starting from the early years of implementation. Secondly, if used as a proxy for income, and in particular corporation income, federal corporation income taxes are suitable as a proxy even in the absence of state corporation income taxes.[52]

Data on internal revenue from 1870-1998 was collected from the "Annual Reports of the Commissioner of Internal Revenue" (henceforth abbreviated as ARCIR), as published by the Office of Internal Revenue, a branch of the Treasury Department. Data from 1999-2020 was collected from the publication "State Government Finances" (henceforth abbreviated as SGF), as published by the Department of Commerce, Bureau of the Census, and as reported within the Census dataset "State_Govt_Fin".

Total personal income tax yield. As mentioned above, data on personal income tax is available from 1915 onwards, and we report as such. With this caveat, the sources for personal income tax are identical to that of internal revenue. Namely, data on personal income tax from 1915-1998 was collected from ARCIR. Data from 1999-2020 is as collected from IRS Statistics on Income Tax Statistics.

Total corporation income tax yield. As mentioned above, data on corporation income tax is available from 1915 onwards, and we report as such. Data on corporation income tax from 1915-1998 was collected from ARCIR. Data from 1999-2020 is as collected from IRS Statistics on Income Tax Statistics.

[52] Note that we report personal income data in 3.7.

Total State government revenue collected. The collection and compilation of state government finances for the 48 states has been a long-running challenge in the literature. This is particularly true pre-1915, where most of the data is spread across individual state-specific government records, in a bevy of state-specific treasurer, auditor and comptroller publications. For the $1870-1915$ period, we build upon the pioneering work of Sylla, Legler and Wallis (1993) and Holt (1970). These works vary in their coverage across states, but are harmonized in Hindman (2010).[53] We use the Hindman (2010) dataset as a starting point for our data on total state government revenue and expenditure, up till 1915.[54] However, we observed in this combined dataset that there are substantial gaps in the revenue series for 24 states, and for the expenditure series in 22 states. We fill these gaps by going into the individual state government reports in the spirit of Sylla, Legler and Wallis (1993).[55] When we go back to the individual state-government reports in an effort to fill these gaps, as much as possible we try to identify the publication used in Sylla, Legler and Wallis (1993) by matching numbers in the existing years reported. We ratio-splice the newly collected data with the corrected Sylla, Legler and Wallis (1993) data as reported in Hindman (2010). Finally, we linearly impute any remaining one and two-year gaps. This mostly applies to states, e.g. Idaho, that have biannual fiscal years. In such cases, we report in the preceding year half of the total revenue for both years, and linearly impute the other year.[56]

From 1916-1919, 1922-1931 and 1937-1950, we digitize data from the publication series "Financial Statistics of States" (henceforth abbreviated as FinStat), which is published by the Department of Commerce, Bureau of the Census.[57] From 1933-1936, we begin with data for 13 states for revenue and 14 states for expenditure, collected from Sylla, Legler and Wallis (2006), as kindly made

[53] We would like to thank Monty Hindman for sharing his datasets with us in personal correspondence. Hindman (2010) takes great care to check and improve the disaggregated figures in Sylla, Legler and Wallis (1993), and is the first to digitize the Holt (1970) data. He combines these two datasets by reporting the respective series whenever available, and taking the average when both are reported.

[54] Holt (1970) collected most of his data while working as a research assistant for Lance Davis, who has published quantitative examinations of the nineteenth-century fiscal conditions with John Legler. This is in contrast to Sylla, Legler and Wallis (1993), who go back to the individual state-government reports.

[55] We note that the Hindman dataset is consistent with the state government finances data on total revenues and total expenditures within Dray, Landais and Stantcheva (2023), who have kindly shared with us their data in personal correspondence.

[56] Across all states, we linearly impute 7 two-year gaps for expenditure and revenue. For these gaps, either states do not publish reports for that period, or we were unable to access scanned copies of the physical documents.

[57] Note that from 1940 onwards, FinStat was published under the name "State Finances". Moreover, in order to ensure consistency in the growth rates, we ratio-splice the pre and post 1915 series. However, data for many states in 1915 are not reported by Sylla, Legler and Wallis (1993). To address this, we collect data from the individual state reports in 1915 for 28 states.

available in ICPSR. From 1951-2008, data is collected from SGF. However, this compiled data presents gaps for all states in 1920, 18 states in 1921, and 35 states in $1933 - 1936$. In order to fill these gaps, we again return to the individual state government reports.

In contrast to internal revenue, State government revenue is collected primarily to be expended by the State government. The main components of revenue include individual and corporation income taxes, property taxes, sales taxes, and insurance trust revenue.[58] Of particular note are property taxes, which are not collected under internal revenue.[59] Since the data is reported at this finer level, we are able to further decompose general revenue into tax revenue and non-tax revenue.[60]

Total state government expenditure. The sources used for data on general expenditure parallel those of general revenue.

Debt. From 1870-1879, we collect data on debt from the publication "Railroads of the United States" by Henry V. Poor, which, for the period in question, included an appendix containing "A Full Analysis of the Debts of the United States, and of the several states.".[61] From 1880-1889, our main source for debt data is the series "American Almanac and Treasury of Facts, Statistical, Financial, and Political", as published by The American News Company.[62] We supplement this data from the publication "Railroads of the United States" for 23 states.[63] The data within was derived in most cases from the officers of the respective states themselves. From 1890-1914, data is taken from SA. We linearly impute any remaining one and two-year gaps, many of which are due to biannual fiscal year reporting cycles. From 1915-1949, data is collected from FinStat. Finally, from 1950-2008, data is collected from SGF.

[58] Note that we exclude the category 'Charges and Miscellaneous' Revenue from the Sylla data for consistency with later years of SGF.

[59] For a detailed treatment of the reasons behind this, see Wallis (2000).

[60] Note that Internal Revenue, in contrast, is comprised of tax revenue.

[61] We note that the publisher H.W. Poor Co evolved to become the Standard & Poor's that we are familiar with today.

[62] We note that some states did not incur debt in the early years. When this is stated in the relevant sources, we code the state's debt as 0 for that year. Similarly, in rare cases states repudiated their debt. For example, Mississippi in 1838, created a debt to the amount of $7,000,000 for the establishment of banks. From the Poor's report in 1877, we learn that Mississippi soon ceased to pay interest on these bonds, and has long since wholly repudiated them. In such cases, we also code the debt as 0.

[63] In particular, CA in 1889, CO in 1886 and 1889, DE in 1889, FL in 1889, GA in 1889, KS in 1889, LA in 1888 and 1889, ME in 1889, MD in 1888, MN in 1888, MO in 1889, NB in 1887 and 1888, NV in 1889, NH in 1888, NJ in 1889, NC in 1887 and 1889, OH in 1888, OR in 1887 and 1889, RI in 1889, TX in 1888, UT in 1889, VT in 1886 and WY in 1888.

Total long-term state government debt. The categorization of the various components of long-term debt have changed multiple times over the years, yet the core components of bonds and debt to trust funds remain the same. Therefore, we are able to retain comparability of the series over the duration of our sample. That being said, we report the changes in categorizations below for reference in comparison with the original documents: From 1880-1936, long-term debt consisted of Funded or Fixed (including bonds), and Floating (including debt to public trust funds) debt. From 1937-1940, it consisted of General obligations (Capital outlays, Funding current expenses, Refunding, Debt to trust funds and others) and Revenue bonds. From 1941-1943, in addition to the previous categories it included Quasi-revenue bonds and Debt serviced by local units. Finally, from 1944-2008, long-term debt corresponds to the sum of Full faith and credit debt, and Non-guaranteed debt.

Total gross state government debt. Gross debt is equal to the sum of long-term debt and "Other" debt not encompassed in long-term debt. From 1880-1913, there was no "Other" debt and therefore long-term debt and gross debt are equal. From 1915-1925, "Other" debt included revenue bonds and notes, warrants and audits and obligations on trust accounts. From 1926-1931, "Other" debt included revenue bonds and notes, warrants and audits, and special assessment bonds and certificates. In 1937, "Other" debt included contingent obligations. From 1938-1939, "Other" debt included contingent obligations and short-term loans. In 1940, "Other" debt included obligations serviced by local units and short-term loans. In 1941, "Other" debt included short-term loans. Finally, from 1942-2008, "Other" debt included short-term debt, and in particular, bond anticipation notes, tax anticipation notes, and warrants.

Total net state government debt. Net debt is defined as long-term debt less total sinking fund assets. Sinking fund assets, in turn, are defined to be funds dedicated to the retirement of long-term debt.

3.7 Labor Market Outcomes

We have gathered 3 data series that deal with labor market outcomes, which are reported annually from 1929 to 2020. These variables provide information on [1] total non-farm employment; [2] total wages & salaries; [3] personal income.

Total non-farm employment. Data from 1870-1930 is compiled from Census, at the Census frequency.[64] Annual data from 1939 onwards is compiled from the BLS series on non-farm employment with a few exceptions. Namely, Illinois, Minnesota and Mississippi are not reported by BLS from 1939-1940, and we digitize the 1940 observations from Census 1940. From 1931 to 1938, we compile data from Wallis (1989), which reports a state-level total non-agricultural employment index from 1929 to 1940.[65] In particular, the data is reported as an index such that 1929 = 100.[66] Since the data is reported as an index, we first construct year-on-year growth rates, and then apply them to the 1939 BLS data to back out non-farm employment in levels from 1931 to 1938.[67] Similarly, we apply the 1930 growth rates from Wallis (1989) to the 1930 Census data in order to back out the 1929 levels. From 1941-1946, we digitize data on non-farm employment in Illinois, Minnesota and Mississippi from the BLS publication "Employment and Payrolls". Whenever possible, we report data collected in December of the year (this applies to the years 1942–1945). In 1946, we were only able to obtain data in November. In order to account for potential seasonal changes from month to month, we weight the November 1946 values by the state-specific ratio of the December 1945/November 1945 values. Similarly, we proxy the employment in December 1941 by using the available January 1942 values, and weighting them by the state-specific ratio of the December 1942/January 1943 values. Though agriculture and therefore farms are economically significant, especially in the early years, the reason that we choose to report non-farm employment is for greater comparability with the later years.

Personal income. 1920 data is taken from Easterlin (1957), who also reports estimates for 1880, 1900 and 1950. For 1880, 1890, 1900 and 1910, data is obtained from Klein (2013), which adopts and revises the methodology used by Easterlin, and updates his 1880-1910 estimates with more recently discovered data. Annual data in 1919 and 1921 is constructed from Leven (1925), which reports data in 1919, 1920 and 1921. We use these three years to construct growth rates in 1920 and 1921, and apply these growth rates to the Easterlin (1957) 1920 data. In fact, the Easterlin (1957) estimates are themselves based off the Leven (1925) estimates, but adjusted to better fit

[64] For the Census data, we construct non-farm employment by taking the total measure of employment, number of gainfully employed, less the total number employed in agriculture.

[65] Note that Wallis (1989) also reports companion indexes for manufacturing employment and non-manufacturing employment.

[66] Note that the index is based in the month of August 1929.

[67] This is with the exception of the Idaho, Michigan, and Minnesota, for which we apply the growth rates to the figures collected from 1940 Census.

the definition of personal income. We proceed in this manner instead of taking the Leven data wholesale in order to take advantage of the updated data by Easterlin. Annual data from 1927-1928 is collected from "Personal Income by States since 1929", a supplement to the Survey of Current Business published by the US Department of Commerce in 1956. Within the report, these two points were constructed using state-level data on income components accounting for three-fifths of personal income, with the other two-fifths constructed using national data in 1927 and 1928, and state shares in 1929. Annual data from 1929-2020 is compiled from the BEA SAINC 4 dataset. In line with the literature, we exclude personal current transfer receipts from total personal income. We note that this is equivalent to taking the sum of "Net earnings by place of residence", and "Dividends, interest, and rent".

3.8 Other State-Level Data

Population. From 1870 to 1890, decennial estimates are taken directly from the Census reports. Annual estimates for 1900—1989 are taken from the Current Population Reports.[68] Annual estimates for 1990-1999 are taken from the US Census Population Division.[69] We perform three imputations on the raw data: 1) *Impute population estimates for July 1, 1970*: Population estimates for 1970 are reported as of April 1 in the Census. These figures are converted to July 1 with the following procedure: [1] compute state-level proportions as of April 1 by dividing each State's population estimate by aggregate estimate; [2] obtain an estimate of US-level resident population for July 1 from the Curent Population Reports Series P-25 No.727; [3] compute the rate of increase from April 1 to July 1 using the US-level estimates; [4] assuming that state-level population proportions on July 1 stay fixed as in April 1, inflate the estimates from April 1 to July 1 by the rate of increase computed in point [3]. 2) *Impute intercensal population estimates from 1870 to*

[68] Annual estimates for 1900—1939 are taken from the Current Population Reports Series P-25 No.139 (issued in June, 1956). Annual estimates for 1940—1949, except for Alaska and Hawaii, are taken from P-25 No.72 (issued in May, 1953), and corresponding estimates for Alaska and Hawaii are from P-25 No.80 (issued in October 1953). Annual estimates over 1950—1959 are taken from P-25 No.304 (issued in April, 1965). Annual estimates for 1960—1969 are taken from P-25 No.460 (issued in June, 1971). Annual estimates for 1970—1979 are taken from P-25 No.957 (issued in October, 1984). Estimates for 1980—1989 are taken from P-25 No.1106 (issued in August, 1996). These sources provide comparable resident population estimates. The P-25 series publications can be found here. A summary of data tables are made available here.

[69] Annual estimates for 1990—1999 are from the Time Series of Intercensal State Population Estimates, Table CO-EST2001-12-00 (released on April 2002 by the US Census Population Division). Estimates for 2000—2009 are taken from Table ST-EST00INT-01 (released on September 2011) from the same source. Then, estimates for 2010—2019 are taken from Table NST-EST2020 (released on July 2021) while estimates for 2020 are from Table NST-EST2022-POP (released on December 2022).

1900: (except for Alaska, Hawaii, and Oklahoma). We perform these imputations following the interpolation method used within the Census technical reports on intercensal estimates. 3) *Impute intercensal population estimates for Oklahoma from 1880 to 1900*: Before the State's formation in November 1907, the estimates reflect the aggregate resident population in Oklahoma and Indian Territories. The intercensal estimates from 1890 to 1900 in Oklahoma follows from point [2] above. Since official population estimate for 1880 is not available, we use a backward extrapolation to impute the population counts from 1881 to 1889. A constant growth rate — calculated by averaging the annual growth rates from 1890 to 1900 — is assumed in the backward extrapolation exercise.

Patents. We report patents as they are a widespread measure of innovation, both historically and in more recent times—see Sokoloff (1988), Moser and Voena (2012), and Sampat (2018). The data concerns patents granted by the United States Patent and Trademark Office (USPTO). A patent inventor is defined as one who contributes to the conception of an invention. A patent assignee is defined as the entity (e.g., company, foundation, partnership, holding company or individual) that is the recipient of a transfer of a patent application, patent, trademark application or trademark registration from its owner of record (assignor). From 1870-2015, data is compiled from Berkes (2018), which has been carefully constructed using the original patent documents.[70] We use the filing year whenever available, and the issue year when not (this accounts for 2.2% of the observations).[71] As noted in Berkes (2018), filing years are arguably a better indicator of when the invention was completed than the issue years. We provide three measures of number of patent counts, two based on the inventor locations and one based on the assignee locations. In our main specification, we assign patent locations based on the location of the first name inventor. We compile data from 1992 to 2020 from the FRED release "U.S. Granted Patents by States, Territories, and Countries", which also assign patent locations based on the location of the first name inventor. While the patent counts in 1992 are quantitatively similar in both datasets, we ratio-splice the two series together using 1992 as a base year. As an alternative specification, we use the 'maximum' number of patents, i.e. for a given patent with inventors who reside in multiple states, we include the single patent in the count for each of the relevant states. We construct a similar measure

[70] We thank Enrico Berkes for kindly sharing with us the CUSP dataset. As noted in Berkes (2018), the number of patents reported in CUSP is similar to that reported in the Petralia, Balland and Rigby (2016) data, apart from the period between World War I and II where CUSP covers more patents.

[71] Additionally, we only keep inventor and assignee data with a state that matches one of the 50 states. This drops 0.7% of patent-inventor observations and 0.7% of patent-assignee observations.

based upon assignee locations. For robustness, we could instead consider the 'minimum' number of patents by only considering patents with inventors/assignees within a unique state.

Sentiments. Data on sentiments is compiled from Van Binsbergen et al. (2024).[72] The data is reported at the state-level, with the earliest observation starting in 1850 and the latest observation ending in 2019. For the purpose of analysis, we consider first differences of the index.

Newspapers. Data on newspapers is primarily compiled from Gentzkow, Shapiro and Sinkinson (2014). The data concerns daily newspapers, and is originally reported at the city-level. We aggregate up to the state-level by taking a simple average. The primary sources of the data are George P. Rowell and Company's (Rowell's) American Newspaper Directory (1869-1876), N.W. Ayer and Son's (Ayer's) American Newspaper Annual (1880-1928), and the Editor and Publisher Yearbook (1932-2004). For 2004-2009, 2013-2014, 2016-2017 for Number of Daily Newspapers Outlets and including 2018 for Newspaper Circulation, we collect the Editor and Publisher Yearbook data as reported in SA.[73] Finally, for the Number of Daily Newspapers in 2018 and for both Number of Daily Newspaper Outlets and Newspaper Circulation in 2020, we collect data from the UNC News Deserts Database.[74] This data is reported on the outlet-city level, and we aggregate up to state-level by taking a simple sum. Finally, we ratio-splice the data using the overlapping year 2018 before merging it with the Editor and Publisher time series from Gentzkow, Shapiro and Sinkinson (2014).

Imports & Exports. We collect data on imports and exports of merchandise at the customs district level, from 1860-2021. Data in 1860-1870,1886-1895,1900-1914,1923-1931,1933-1946,1980,1983 and 1988 are digitized from SA. Data in 1871-1885 are digitized from "Quarterly Reports of the Chief of the Bureau of Statistics Showing the Imports and Exports of the United States", as published by the Department of the Treasury, Bureau of Statistics. Data in 1898-1899,1915-1917,1918-1920,1932-1933,1950-1959 are digitized from the publications "Foreign Commerce of the United States", "Monthly Summary of Foreign Commerce of the United States" and "Quarterly Summary

[72] We thank Varun Sharma and coauthors for kindly sharing their dataset.

[73] We note that the intervening year data is collected within the Editor and Publisher Yearbook, but is unfortunately not publicly available. We correct for an error in the 2014 and 2016 data as reported in SA 2017 and 2018 respectively which lists North Dakota (ND) twice in place of North Carolina (NC).

[74] We would like to thank Elizabeth Thompson from the UNC Hussman School of Journalism and Media for kindly sharing the data with us.

of Foreign Commerce of the United States", as published by the United States Bureau of Foreign and Domestic Commerce. Data in 1960-1965 are digitized from "Foreign Commerce and Navigation of the United States", as published by United States Department of the Treasury, Bureau of Statistics. Data in 1966-1988 are digitized from "Foreign Trade: Highlights of U.S. Export and Import Trade", as published by the United States Department of Commerce. Finally, data from 1989-2021 are collected from Schott (2008). The states that each customs district belongs to are only sometimes reported in the raw tables. Whenever they are not reported, we manually assign the customs district to the corresponding states. Finally, we aggregate to the state level by taking a simple sum across customs district for each state.[75]

House & Rent Prices. Data on house prices from 1890-1975 and rent prices from 1890-2006 is collected from Lyons et al. (2024), which is reported on the city level. We aggregate the data to the state-level by first computing the growth rates of each index at the city-level, then taking the simple average across cities in a given state. This approach yields data for 23 states. We collect data on house prices from 1976-2021 from the FHFA House Price Index. We use the non-seasonally adjusted quarterly data, and aggregate to annual data by taking a simple average, before taking growth rates. We report the growth rates from the FHFA House Price Index from 1976-2021.

GDP. Data on state-level nominal GDP from 1963-2022 is collected from BEA, while data on national nominal GDP from 1790-2022 is collected from Williamson (2024).[76] Data on state-level real GDP from 1963-2022 is collected from BEA. In order to obtain a consistent dataset for state-level real GDP in chained 2012 dollars, we proceed in three steps. First, we obtain the latest state-level real GDP estimates from 1997-2022 (chained 2012 dollars), along with the latest state-level real GDP estimates from 1977-1997 (chained 1997 dollars). Second, in order to make the two datasets comparable, we adjust the 1977-1997 data by the ratio of data in the overlapping year 1997. Third, since there is no available real data before 1977, we adjust the state-level nominal data from 1963-1977 by the US-level GDP deflator, and reweight the estimates based upon the shares of the overlapping year 1977, such that the 1977 data in both the above samples is the same. To do so,

[75] Note that customs district are occasionally reported together under multiple states, for example, Montana and Idaho in 1877. In such cases, we split the total value between the relevant states, using population weights in the corresponding year.

[76] Note that from 1929-2022, Measuring Worth draws upon the BEA estimates.

we use the latest BEA national GDP deflator estimates as retrieved from FRED (adjusted to 2012 base year). These estimates span from 1929 - 2023. Finally, to obtain a national real GDP data series from 1790-2022, we harness the long-running estimates of real GDP from Williamson (2024), updating their 1929-2022 national price deflators to the latest BEA estimates and reweighting to base year 2012.

Banking. We take data on bank assets, bank deposits, bank loans, bank capital and bank liabilities covering the period 1863-2020 from Hoon, Liu, Payne, Müller and Zheng (2025). The data concerns National and State Banks. From 1863 - 1960, the data has been digitized from the Officer of the Comptroller of the Currency's (OCC) Annual Reports. From 1961 - 1980, the main source of the data is from the Federal Deposit Insurance Corporation (FDIC) database. A potential challenge through the 1980s comes about due to a period of Financial Deregulation, where States began to allow interstate branch banking. In particular, the FDIC data allocates banks to states by headquarters, regardless of branch location. From 1981 onwards, we draw on the Federal Financial Institutions Examination Council (FFIEC)'s Consolidated Reports of Condition and Income, better known as "Call Reports". Since we are interested in the state-level data, we use the Summary of Deposits branch-level data to construct bank-state-year shares, then apply them to the Call Reports data to obtain state-level values. Finally, we splice the adjusted data with the pre-1961 OCC data. More details can be found in our companion paper on US Financial Crises (Hoon, Liu, Payne, Müller and Zheng, 2025).

4 Variable Availability By State

To showcase the availability of state-level economic variables we have compiled, Section 5 contains one plot for each state. Every page shows the exact availability of each variable in our dataset to give a sense of the broad scope of the digitization effort that went into constructing these time series.

5 Acknowledgments

This project would not have been possible without the generous help of dozens of academics and subject matter experts that generously provided us data or helped us to identify and understand many of the historical data sources. In this section, we provide a full list of acknowledgments to show our gratitude.

We thank the staff of Purdue Libraries and in particular Haley Bond for assistance related to business failure data. We thank the staff of Indiana University Bloomington Libraries and in particular Craig J. Clark for assistance related to business failure data. We thank Kathleen Bonk from the New York State Museum for assistance with New York Petroleum data.

We thank Kristine Sheaffer (USGS) for assistance with Gold and Silver statistics. We thank Monika Ghimire (USDA) for assistance with the ERS and Cotton data, and Chris Singh (USDA) for assistance with Sweet Potato data.

We thank Elizabeth Thompson (UNC Hussman) for kindly sharing the UNC News Desert Newspaper data.

We thank Christine Hartley and the team from the U.S. Census Bureau (Population Division) for their feedback on our imputation of state-level intercensal population estimates prior to 1900.

Lastly, we thank Lang Yang, Zihui Wang, Jacob Wang, Yuanyuan Dong, and Jia Yeong Ng for excellent research assistance. We thank Zhihan Liu for assistance and comments related to business failure data.

References

Berkes, Enrico. 2018. "Comprehensive Universe of Us Patents (Cusp): Data and Facts." *Unpublished, Ohio State University* .

Björck, Åke. 2024. *Numerical Methods for Least Squares Problems: Second Edition*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Boot, Johannes Cornelius Gerardus, Walter Feibes and Johannes Hubertus Cornelius Lisman. 1967. "Further methods of derivation of quarterly figures from annual data." *Applied Statistics* pp. 65–75.

Census. 1963. Annual Survey of Manufactures Estimates Compared with 1963 Census of Manufactures Totals. Special report U.S. Department of Commerce, Bureau of the Census.

Chow, Gregory C and Anloh Lin. 1971. "Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series." *The Review of Economics and Statistics* pp. 372–375.

Craig, James R and J.Donald Rimstidt. 1998. "Gold production history of the United States." *Ore Geology Reviews* 13(6):407–464.

Denton, Frank T. 1971. "Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization." *Journal of the American Statistical Association* 66(333):99–102.

Dray, Sacha, Camille Landais and Stefanie Stantcheva. 2023. Wealth and Property Taxation in the United States. Technical report National Bureau of Economic Research.

Easterlin, Richard A. 1957. "State Income Estimates." *Population redistribution and economic growth United States 1870–1950, Vol. 1: Methodological considerations and reference tables.* 1957b.

Energy Information Administration. 1976. Petroleum Statement: Crude Petroleum, Petroleum Products, and Natural Gas Liquids. Energy data reports Department of Energy.

Fernández, R B. 1981. "A Methodological Note on the Estimation of Time Series." *The Review of Economics and Statistics* pp. 471–476.

Fritsch, Frederick N and Ralph E Carlson. 1980. "Monotone Piecewise Cubic Interpolation." *SIAM Journal on Numerical Analysis* 17(2):238–246.

Gentzkow, Matthew, Jesse M. Shapiro and Michael Sinkinson. 2014. "United States Newspaper Panel, 1869-2004.".

Hansen, M. E. and B. A. Hansen. 2012. "Crisis and Bankruptcy: The Mediating Role of State Law, 1920–1932." *The Journal of Economic History* 72(2):448–468.

Hansen, M. E. and N. L. Ziebarth. 2017. "Credit Relationships and Business Bankruptcy during the Great Depression." *American Economic Journal: Macroeconomics* 9(2):228–255.

Hansen, Mary Eschelbach, Matthew Davis and Megan Fasules. 2016. United States Bankruptcy Statistics by District. 1899-2007. Technical report Inter-University Consortium for Political and Social Research.

Herfindahl, Orris C. 1966. *Output, Employment, and Productivity in the United States after 1800.* NBER Chapters National Bureau of Economic Research, Inc chapter Development of the Major Metal Mining Industries in the United States from 1839 to 1909.

Hindman, Monty. 2010. The Rise and Fall of Wealth Taxation: An Inquiry Into the Fiscal History of the American States PhD thesis University of Michigan.

Holt, Charles Frank. 1970. *The role of state government in the nineteenth-century American economy, 1820-1902: a quantitative study.* Purdue University.

Hoon, Joseph, Chang Liu, Jonathan Payne, Karsten Müller and Zhongxi Zheng. 2025. "The Costs of Financial Crises in the United States." *Working Paper* .

Hoon, Joseph, Chang Liu, Karsten Müller and Zhongxi Zheng. 2025. "U.S. State-Level Business Cycles Since the Civil War." *Working Paper* .

Jaworski, Taylor. 2017. "World War II and the Industrialization of the American South." *The Journal of Economic History* 77(4):1048–1082.

Klein, Alexander. 2013. New State-Level Estimates of Personal Income in the United States, 1880–1910. In *Research in economic history*. Emerald Group Publishing Limited pp. 191–255.

Leven, Maurice. 1925. "Income in the Various States: Its Sources and Distribution, 1919, 1920, and 1921". In *Income in the Various States: Its Sources and Distribution, 1919, 1920, and 1921*. NBER pp. 41–50.

Litterman, R B. 1983. "A Random Walk, Markov Model for the Distribution of Time Series." *Journal of Business & Economic Statistics* pp. 169–173.

Lucier, Gary. 1986. *Farm Income Data: A Historical Perspective*. US Department of Agriculture, Economic Research Service.

Lyons, Ronan C, Allison Shertzer, Rowena Gray and David N Agorastos. 2024. The Price of Housing in the United States, 1890-2006. Technical report National Bureau of Economic Research.

Merrill, Charles White. 1930. Summarized Data of Silver Production. Economic Paper 8 U.S. Department of Commerce Bureau of Mines.

Moser, Petra and Alessandra Voena. 2012. "Compulsory Licensing: Evidence from the Trading with the Enemy Act." *American Economic Review* 102(1):396–427.

Naples, M. I. and A. Arifaj. 1997. "The Rise in US Business Failures: Correcting the 1984 Data Discontinuity." *Contributions to Political Economy* 16(1):49–59.

Petralia, Sergio, Pierre-Alexandre Balland and David Rigby. 2016. "HistPat Dataset.".

Richardson, Gary and Michael Gou. 2011*a*. Business Failures by Industry in the United States, 1895 to 1939: A Statistical History. Working Paper 16872 National Bureau of Economic Research.

Richardson, Gary and Michael Gou. 2011*b*. "Business Failures by Industry in the United States, 1895 to 1939: A Statistical History." *NBER Working Papers* .

Sampat, Bhaven N. 2018. A Survey of Empirical Evidence on Patents and Innovation. Working Paper 25383 National Bureau of Economic Research.

Schott, Peter K. 2008. "The Relative Sophistication of Chinese Exports." *Economic Policy* 23(53):6–49.

Simpson, Paul B. and Paul S. Anderson. 1957. "Liabilities of Business Failures as a Business Indicator." *The Review of Economics and Statistics* 39(2):193–199.

Sokoloff, Kenneth L. 1988. Inventive Activity in Early Industrial America: Evidence From Patent Records, 1790 - 1846. Working Paper 2707 National Bureau of Economic Research.

Strauss, Frederick and Louis Hyman Bean. 1940. *Gross Farm Income and Indices of Farm Production and Prices in the United States, 1869-1937*. US Department of Agriculture.

Sylla, Richard E., John B. Legler and John Wallis. 1993. "Sources and Uses of Funds in State and Local Governments, 1790-1915: [United States].".

Sylla, Richard E., John B. Legler and John Wallis. 2006. "State and Local Government [United States]: Sources and Uses of Funds, State Financial Statistics, 1933-1937.".

Towne, Marvin and Wayne Rasmussen. 1960. *Farm Gross Product and Gross Investment in the Nineteenth Century*. Princeton University Press pp. 255–316.

Van Binsbergen, Jules H, Svetlana Bryzgalova, Mayukh Mukhopadhyay and Varun Sharma. 2024. (Almost) 200 Years of News-Based Economic Sentiment. Technical report National Bureau of Economic Research.

Wallis, John Joseph. 1989. "Employment in the Great Depression: New data and hypotheses." *Explorations in Economic History* 26(1):45–72.

Wallis, John Joseph. 2000. "American Government Finance in the Long Run: 1790 to 1990." *Journal of Economic Perspectives* 14(1):61–82.

Weiss, Thomas. 1993. "Long-Term Changes in Us Agricultural Output per Worker, 1800-1900." *Economic History Review* pp. 324–341.

Williamson, Samuel H. 2024. What Was the U.S. GDP Then? Technical report Measuring Worth.

Figure 1: Examples of Digitized Sources

STATE DEBTS, VALUATION, AND TAXES.

THE following statistics of the finances of the thirty-eight States in the Union have been derived in most cases from the officers of the States themselves.

STATES.	Date of Statement.	AMOUNT OF STATE DEBT. Funded.	Unfunded.	Amount Raised by Taxation Last Year.	AMOUNT OF TAXABLE PROPERTY AS ASSESSED. Real.	Personal.	State Tax Per Cent on $100.
		$	$	$	$	$	Cts.
Alabama....	..... 1876.	14,061,670	.....	.....	83,851,252	76,200,000	75
Arkansas..	Sept. 30, 1877.	4,153,035	13,967,012	457,450	61,960,452	32,692,425	60
California..	June 30, 1877.	3,411,000	.....	4,105,884	454,641,311	140,431,866	73½
Colorado...	Nov. 1, 1876.	None.	50,000	74,000	25,584,669	18,545,536	15
Connect't..	Dec. 1, 1876.	5,014,500	.....	705,024	238,027,032	106,379,945	15
Delaware...	.....	1,000,000 (?)	.....	.....	.....	.....	50
Florida.....	Jan. 1, 1877.	1,259,600	43,392	225,000	19,713,462	10,197,991	70
Georgia	Jan., 1877.	11,135,500	.....	.....	146,041,809	99,811,941	50
Illinois.....	Jan. 1, 1876.	796,330	None.	2,640,025	931,199,308	197,291,421	36
Indiana.....	Nov. 1, 1877.	1,097,755	.....	1,395,484	638,246,860	222,362,781	13
Iowa.....	Oct. 30, 1877.	545,435	341,000	(?) 750,000	324,698,364	79,971,680	20
Kansas.....	Nov. 1, 1877.	1,235,975	.....	714,549	94,586,003	29,246,313	55
Kentucky..	Oct. 10, 1876.	2,477,000	183,394	1,586,138	211,508,998	396,584,486	40
Louisiana..	May 31, 1876.	9,318,343	2,548,812	2,473,629	139,220,457	35,485,337	14½
Maine	Jan. 1, 1877.	5,920,400	.....	675,173	224,579,569 (Real & Personal.)		30
Maryland ..	..... 1877.	10,296,522*	.....	.....	547,044,271 (Real & Personal.)		17½
Massac'ts ..	Jan. 1, 1877.	33,550,464†	17,072	1,800,000	1,191,499,228	508,965,487	10

(a) American Almanac: State Debt

COMMERCIAL FAILURES IN 1895.

STATES.	COMMERCIAL FAILURES. TOTAL, 1895. No.	Assets.	Liabilities.	TOTAL, 1894. No.	Liabilities.	CLASSIFIED FAILURES, 1895. MANUFACTURING. No.	Liabilities.	TRADING. No.	Liabilities.	OTHER COM'L. No.	Liabilities.	BANKING. No.	Liabilities.
Maine.....	188	$496,156	$1,257,858	251	$2,449,210	35	$482,200	152	$774,158	1	$1,500	1	$80,000
New Hampshire.	58	213,721	405,644	47	326,646	11	75,800	47	329,844	..	.....	..	.....
Vermont	36	64,400	145,300	33	315,096	6	34,000	30	111,300	..	.....	..	.....
Massachusetts...	567	4,342,003	10,942,638	836	16,467,631	288	5,376,080	270	4,849,891	9	716,667	1	45,000
Connecticut.....	254	1,786,236	2,442,980	253	1,821,143	73	1,704,110	177	728,870	4	10,000	2	526,000
Rhode Island....	202	573,925	3,771,397	187	1,480,566	46	2,866,511	137	855,486	19	49,400	1	1,166,526
New England.	1,305	$7,476,441	$18,965,817	1,607	$22,860,292	459	$10,538,701	813	$7,649,549	33	$777,567	5	$1,817,526
" 1894	1,607	9,889,410	22,860,292			452	10,499,011	1,140	12,014,956	15	346,325	1	125,000
New York	1,940	$23,033,614	$45,225,534	1,976	$36,858,225	560	$25,985,159	1,344	$17,616,587	36	$1,623,788	4	$2,647,179
New Jersey	182	2,655,203	3,612,488	212	3,351,766	62	2,462,601	113	858,789	7	291,098	..	.....
Pennsylvania ...	1,349	7,494,071	11,739,947	1,433	15,685,058	368	4,566,682	975	7,121,065	6	52,200	6	745,434
Middle.....	3,471	$33,182,888	$60,577,969	3,621	$55,895,049	990	$33,014,442	2,432	$25,596,441	49	$1,967,086	10	$3,392,613
" 1894	3,621	31,337,202	55,895,049			1,150	26,415,912	2,399	26,415,735	72	3,065,402	15	7,452,724
Maryland	299	$1,825,784	$3,279,124	239	$2,971,319	86	$1,467,362	205	$1,671,706	8	$140,056	..	.....
Delaware	68	194,250	460,100	61	936,770	10	170,500	58	289,600	..	.....	..	.....
Dist. Columbia..	58	567,386	713,609	52	854,855	7	92,460	50	585,824	1	35,325	1	$16,884
Virginia	305	1,733,364	2,929,414	267	2,026,042	27	446,000	275	2,151,414	3	332,000	2	1,198,402
West Virginia...	69	402,572	691,324	100	532,279	17	263,533	51	426,791	1	1,000	..	.....
North Carolina..	103	799,317	1,231,747	139	1,901,810	5	72,000	98	1,159,747	..	.....	..	.....
South Carolina..	102	1,097,763	1,263,903	97	2,121,815	8	358,453	94	905,450	..	.....	..	.....
Florida	136	1,933,400	1,741,550	46	380,650	4	34,000	131	1,704,050	1	3,500	3	450,000
Georgia	214	2,093,006	3,049,383	347	4,756,118	21	1,191,750	188	1,773,833	5	83,800	2	365,000
Alabama.....	140	757,000	1,320,250	190	2,944,309	11	250,500	129	1,069,750	..	.....	..	.....
Mississippi.....	115	829,720	935,340	171	1,357,699	1	4,000	113	930,340	1	1,000	..	.....
Louisiana.....	199	2,743,733	2,876,081	230	1,897,799	17	293,348	181	2,570,733	1	12,000	2	236,282
Tennessee.....	273	2,416,622	2,646,632	347	3,141,249	23	632,293	247	1,812,569	3	201,770	1	107,800
Kentucky.....	274	2,310,004	3,042,045	339	5,407,830	38	1,859,961	235	1,181,384	1	700	..	.....
South.....	2,355	$19,703,921	$26,180,502	2,625	$31,230,544	244	$7,136,160	2,055	$18,233,191	25	$811,151	11	$2,374,368
" 1894	2,625	25,454,259	31,230,544			293	9,860,361	2,304	19,450,990	28	1,919,193	12	935,254

(b) Dun's Review: Business Failures

Table 2: Description of Raw Data

Variable	Code	Period	Frequency	Sample	Main Sources
Total Value of Agricultural Production ($)	PRODAR0	1870–2021	Annual	All States	USDA
Farm Value per Acre ($)	FARMPA2	1870–1910 1910–2020	Decennial Annual	All States All States	Quickstats
Cash receipts from Crops Sold ($)	PRODAR1	1870–2021	Annual	All States	USDA ERS FIWS/Strauss & Bean/Quickstats
Cash receipts from Animal Products Sold ($)	PRODAR2	1870–2021	Annual	All States	USDA ERS FIWS /Strauss & Bean/Quickstats
Cash Receipts from Forest Products Sold ($)	PRODAR3	1870–2021	Annual	All States	USDA ERS FIWS/Strauss & Bean/Quickstats
Number of Farm Operations	FARMFOP	1870–1910 1910–2021	Decennial Annual	All States All States	NASS census Quickstats
Value of Oats Crop ($)	FARMCR0	1870–2021	Annual	All Producing States	USDA/Quickstats
Value of Wheat Crop ($)	FARMCR1	1870–2021	Annual	All Producing States	USDA/Quickstats
Value of Corn Crop ($)	FARMCR2	1870–2021	Annual	All Producing States	USDA/Quickstats
Value of Barley Crop ($)	FARMCR3	1870–2021	Annual	All Producing States	USDA/Quickstats
Value of Tobacco Crop ($)	FARMCR4	1870–2021	Annual	All Producing States	USDA/Quickstats
Value of Potato Crop ($)	FARMCR5	1870–2021	Annual	All Producing States	USDA/Quickstats

Continued on next page

Variable	Code	Period	Frequency	Sample	Main Sources
Value of Sweet Potato Crop ($)	FARMCR6	1870–2021	Annual	All Producing States	USDA/Quickstats
Value of Cotton Crop ($)	FARMCR6	1870–2021	Annual	All Producing States	Strauss & Bean/USDA/Quickstats
Value of Cottonseed Crop ($)	FARMCR7	1870–2021	Annual	All Producing States	Strauss & Bean/USDA/Quickstats
Value of Rye Crop ($)	FARMCR8	1870–2021	Annual	All Producing States	Strauss & Bean/USDA/Quickstats
Value of Hay Crop ($)	FARMCR9	1909–2021	Annual	All Producing States	USDA/Quickstats
Receipts of Cattle Livestock ($)	FARMLS0	1870–2021	Annual	All Producing States	USDA
Receipts of Hogs Livestock ($)	FARMLS1	1870–2021	Annual	All Producing States	USDA
Value of Sheep Livestock ($)	FARMLS2	1870–1960	Annual	All Producing States	USDA
Value of Horses Livestock ($)	FARMLS3	1870–1960	Annual	All Producing States	USDA
Value of Mules Livestock ($)	FARMLS3	1870–1960	Annual	All Producing States	USDA
Value of Lumber	FARMFS0	1869–1899	Decennial	All States	USDA (Steer 1948)
		1904–1945	Annual	All States	
Manufacturing Establishment Counts	MANUEST	1870–1900	Decennial	All States	Census/ASM
		1905–1915	5-yearly	All States	
		1919–1939	2-yearly	All States	CBP
		1946–2021	Annual	All States	
Manufacturing Employee Counts	MANUEMP	1870–1900	Decennial	All States	Census
		1905–1915	5-yearly	All States	
		1919–2021	Annual	All States	ASM/Wallis/EMPP

Continued on next page

Variable	Code	Period	Frequency	Sample	Main Sources
Manufacturing Wages & Salaries ($)	MANUPAY	1870–1900	Decennial	All States	Census/ASM
		1905–1915	5-yearly	All States	
		1919–1939	Annual	All States	
		1947–2021	Annual	All States	
Manufacturing Value Added ($)	MANUVAL	1870–1900	Decennial	All States	Census/ASM
		1905–1915	5-yearly	All States	
		1919–1939	Annual	All States	
		1947–2021	Annual	All States	
Total value of Mineral Production ($)	MINEVAL	1880–2021	Annual	All Producing States	MinRes/MinYears/EIA
Quantity of Petroleum Produced (barrels)	MINEPET	1870–2021	Annual	All Producing States	MinYears/EIA
Quantity of Coal Produced (tons)	MINECOAL	1870–2021	Annual	All Producing States	MinYears /USCOAL
Quantity of Gold Produced (troy ounces)	MINEGOLD	1870–2021	Annual	All Producing States	MinYears/Craig/SA
Quantity of Silver Produced (fine ounces)	MINESIL	1870–2021	Annual	All Producing States	MinYears/Merrill/SA
Quantity of Pig Iron Produced (tons)	MINEPIRON	1878–2021	Annual	All Producing States	MinYears/SA
Mileage Owned of Railroad Tracks (mi)	TPTRAILO	1870–1973	Annual	All States	Poors/SA
Rural and Municipal Road Mileage (mi)	TPTRROAD	1904–1914	Decennial	All States	SA
		1921–1930	Annual	All States	Highstat
		1942–2000	Annual	All States	
State Highway Mileage (mi)	TPTHIGHW	1923–1941	Annual	All States	SA
		1953–2000	Annual	All States	Highstat

Continued on next page

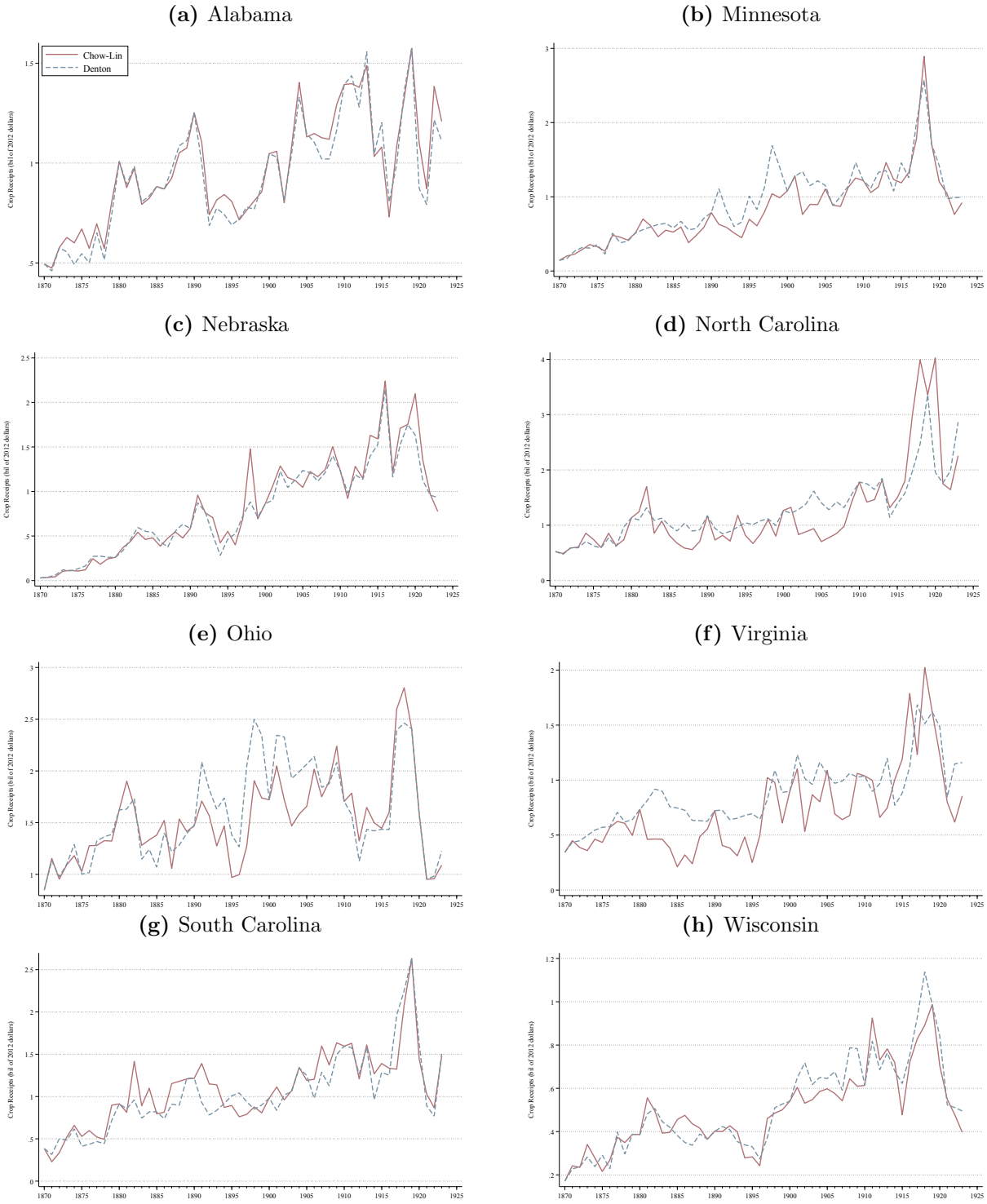
Variable	Code	Period	Frequency	Sample	Main Sources
Number of Registered Motor Vehicles	TPTHMV	1900–2021	Annual	All States	FHWA/Highstat
Automobile Revenues ($)	TPTAREV	1913–2000	Annual	All States	FHWA/Highstat
No. of Business Failures	BIZFAIL	1885–2021	Annual	All States	SA/Hansen/US Courts
Liabilities of Failed Firms ($)	BIZFLIA	1885–1983	Annual	All States	SA
No. of Business Concerns	BIZCONC	1885–1968	Annual	All States	SA
No. of Bankruptcies Commenced	BIZBKCM	1900–2021	Annual	All States	Hansen/US Courts
No. of Bankruptcies Closed	BIZBKCM	1900–2021	Annual	All States	Hansen/US Courts
Federal Govt Internal Revenue Collected ($)	GFIR	1870–2021	Annual	All States	ARCIR/SOI/SGF
Personal Income Tax Yield ($)	GFPIT	1915–2021	Annual	All States	ARCIR/SOI
Corporate Income Tax Yield ($)	GFCIT	1915–2021	Annual	All States	ARCIR/SOI
State Govt General Revenue Collected ($)	GFGR	1870–2021	Annual	All States	Hindman/FinStat/Ind State Rpts
State Govt General Expenditure ($)	GFEXP	1870–2021	Annual	All States	Hindman/FinStat/Ind State Rpts
State Govt Gross Debt ($)	GFGD	1870–2021	Annual	All States	Poors/AA/FinStat/SGF
State Govt Long-Term Debt ($)	GFLTD	1880–2008	Annual	All States	AA/FinStat/SGF
State Govt Net Debt ($)	GFND	1880–2008	Annual	All States	AA/FinStat/SGF
Total Non-Farm Employment	LMNFE	1870–2020	Decennial	All States	Census/BLS
		1929–2021	Annual	All States	BLS/Wallis
Personal Income ($)	LMNPI	1880–1920	Decennial	All States	Klein/Easterlin
		1919–1921 & 1927–2021	Annual	All States	Leven/SCB/BLS/ARCIR
Population	POPN	1870–2020	Annual	All States	Census/Census CPT

Continued on next page

Variable	Code	Period	Frequency	Sample	Main Sources
Value of Imports of Merchandise ($)	IEIMP	1869–1948,1950–1952,1954–1981,1983–2021	Annual	43 States (Cust District level)	FC/QSIE/FT/SA/Schott
Value of Exports of Merchandise ($)	IEEXP	1869–1948,1950–1952,1954–1981,1983–2021	Annual	40 States (Cust District level)	FC/QSIE/FT/SA/Schott
Number of Patents by Inventors (First Name)	PATAFN	1870–2020	Annual	All States	CUSP/FRED
Number of Patents by Inventors (Max)	PATAM	1870–2014	Annual	All States	CUSP
Number of Patents by Assignees (Max)	PATAM	1870–2014	Annual	All States	CUSP
Sentiments	SENT	1850–2019	Annual	All States	Binsbergen
House Prices	HPC	1890–2021	Annual	All States	Lyons/FHFA HPI
Rent Prices	RPC	1890–2006	Annual	23 States (City-lvl)	Lyons/FHFA HPI
Total Circulation of Newspapers	NPC	1869–2004	4-Yearly	All States	Gentzkow
		2004–2020	Annual	All States	SA/UNC
No. of Newspapers operating daily	NPO	1869–2004	4-Yearly	All States	Gentzkow
		2004–2020	Annual	All States	SA/UNC

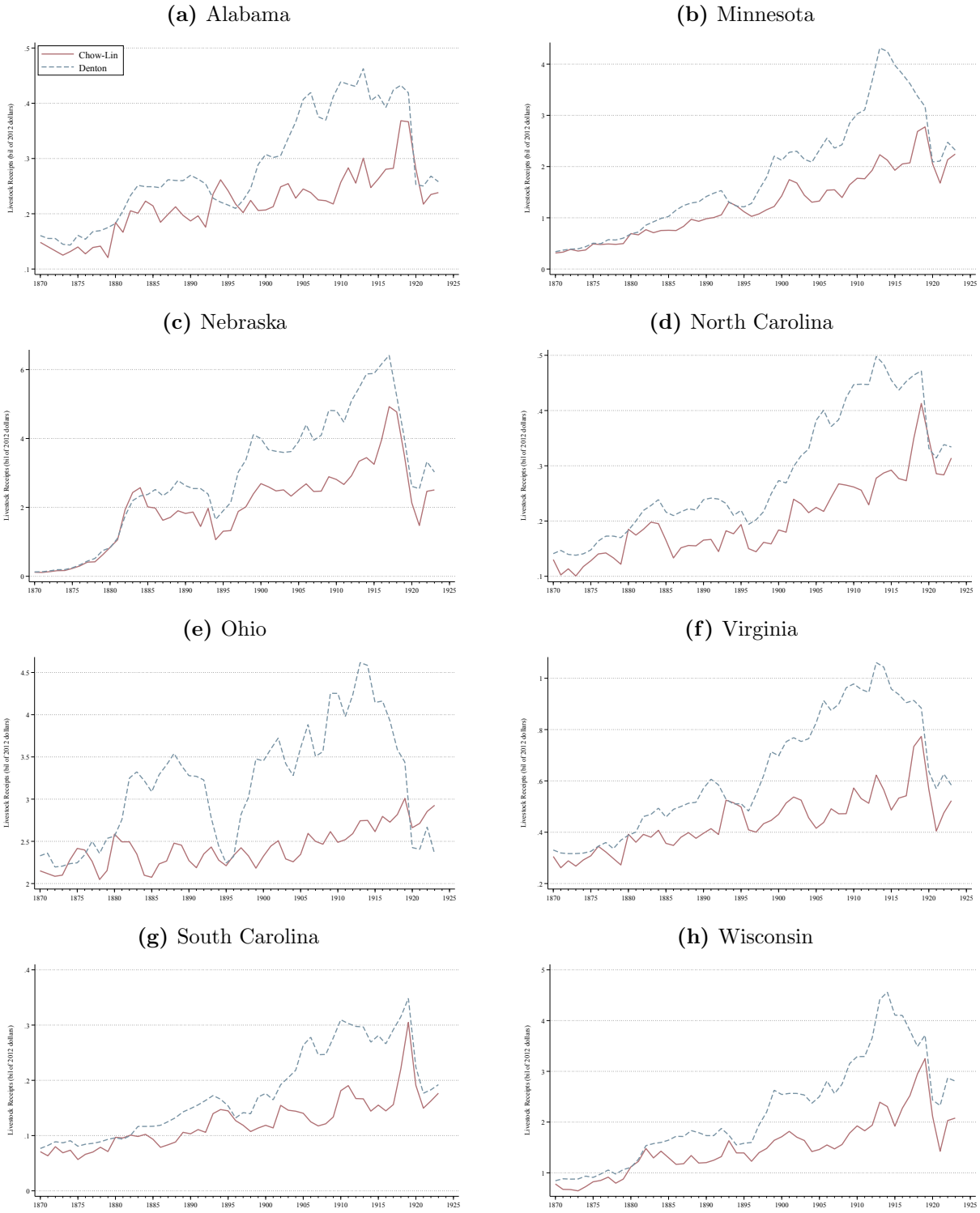
Notes: The first column reports the variable names. The second column provides the corresponding code labels used in our data sets. The third, fourth, and fifth columns describe the available period, frequency, and sample coverage across states, respectively. Links to main data sources are provided in the last column.

Figure 2: Crop Receipts Imputation



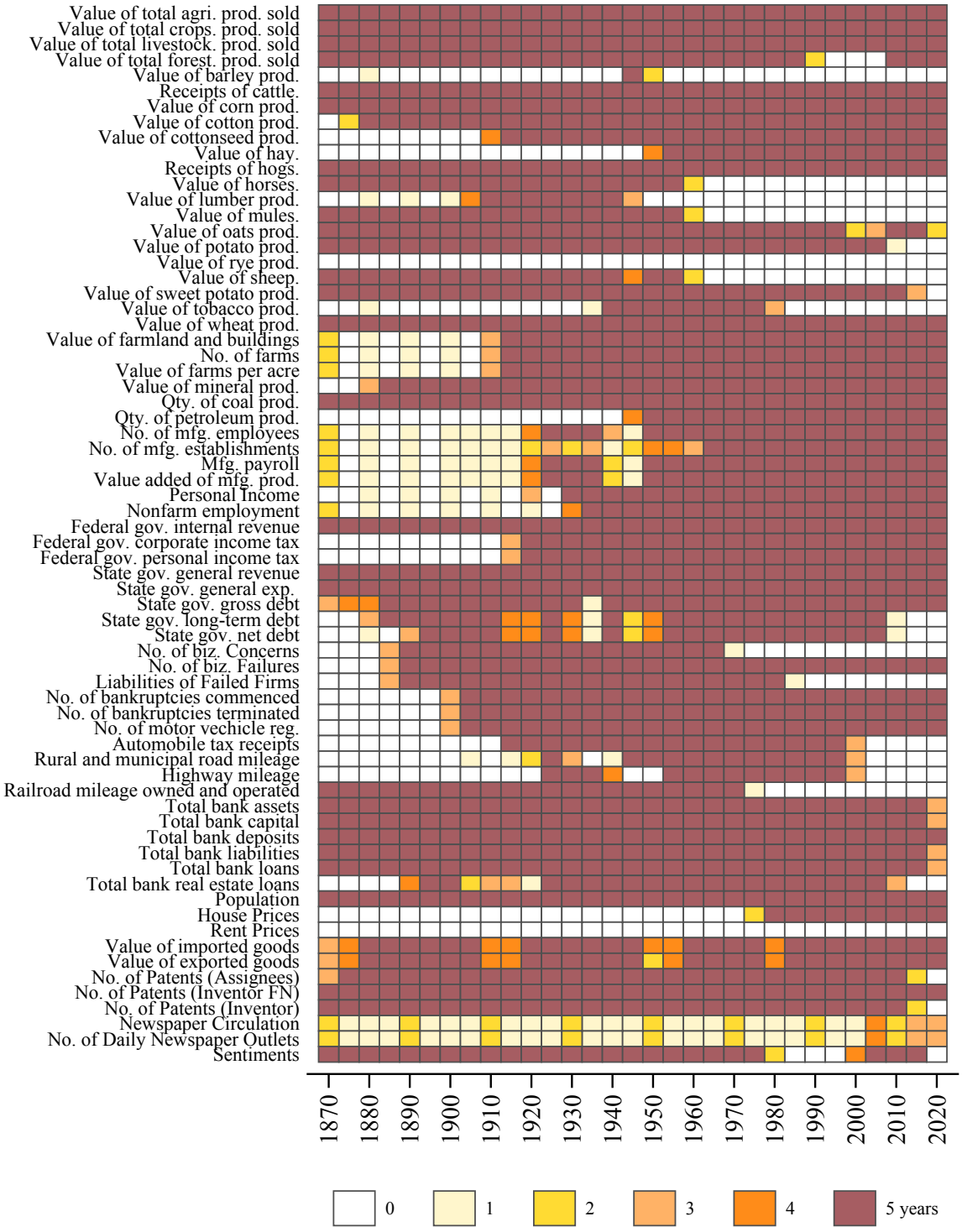
Notes: These plots display a comparison of our imputed crop receipts from 1870 until 1923 using the Chow-Lin and Denton methods. The receipts are denominated in billions of 2012 dollars.

Figure 3: Livestock Receipts Imputation



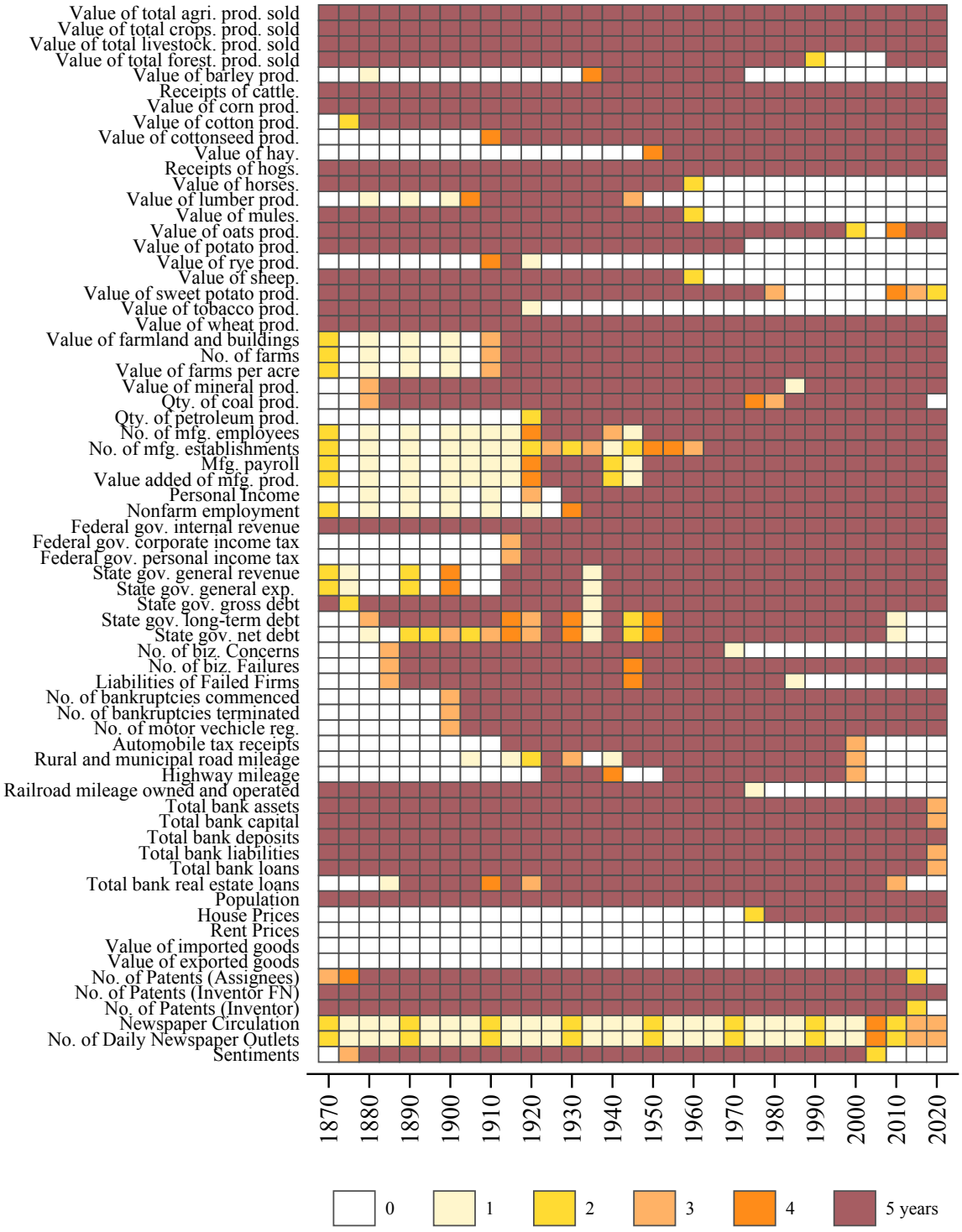
Notes: These plots display a comparison of our imputed livestock receipts from 1870 until 1923 using the Chow-Lin and Denton methods. The receipts are denominated in billions of 2012 dollars.

Figure 4: Availability of Variables – Alabama



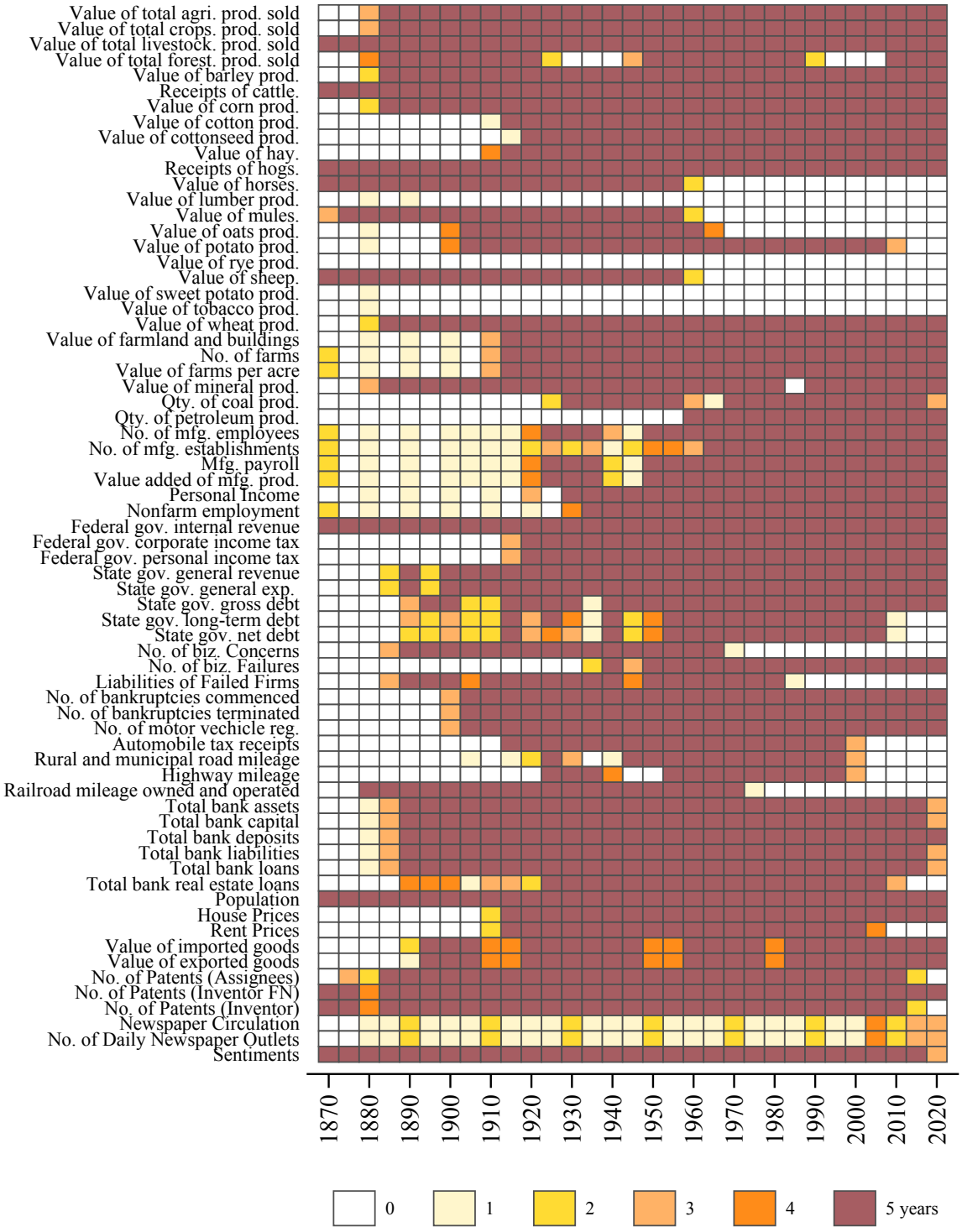
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 5: Availability of Variables – Arkansas



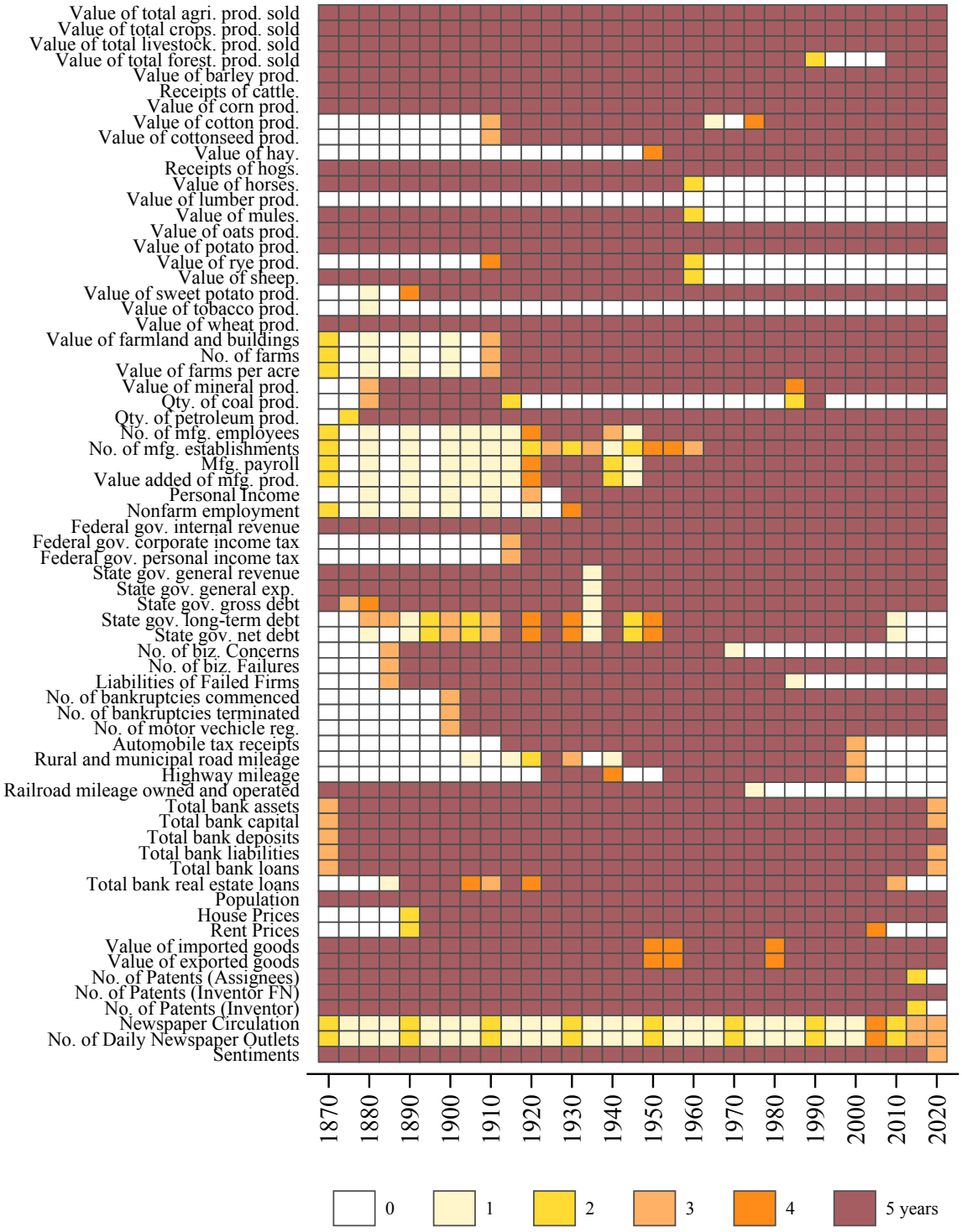
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 6: Availability of Variables – Arizona



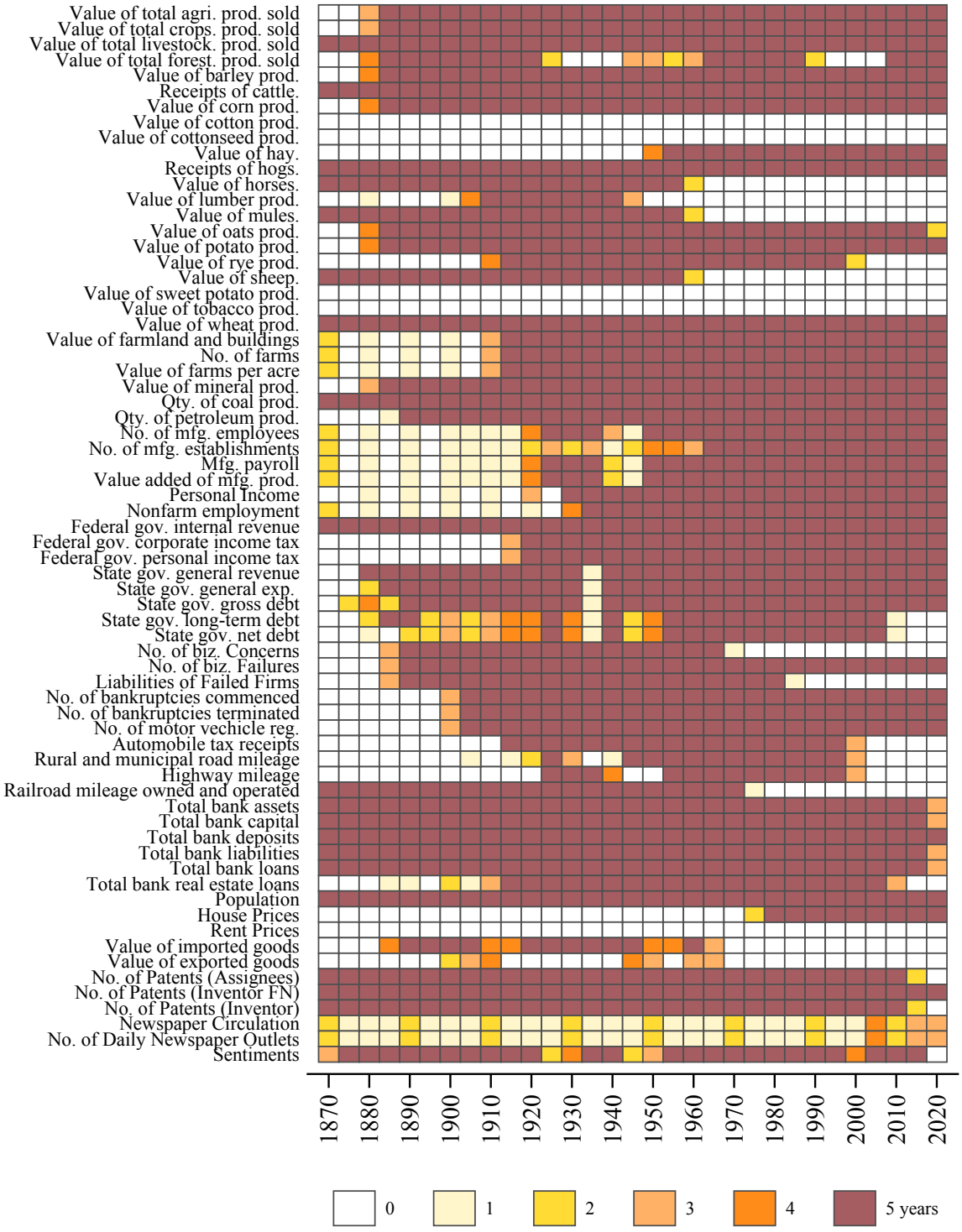
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 7: Availability of Variables – California



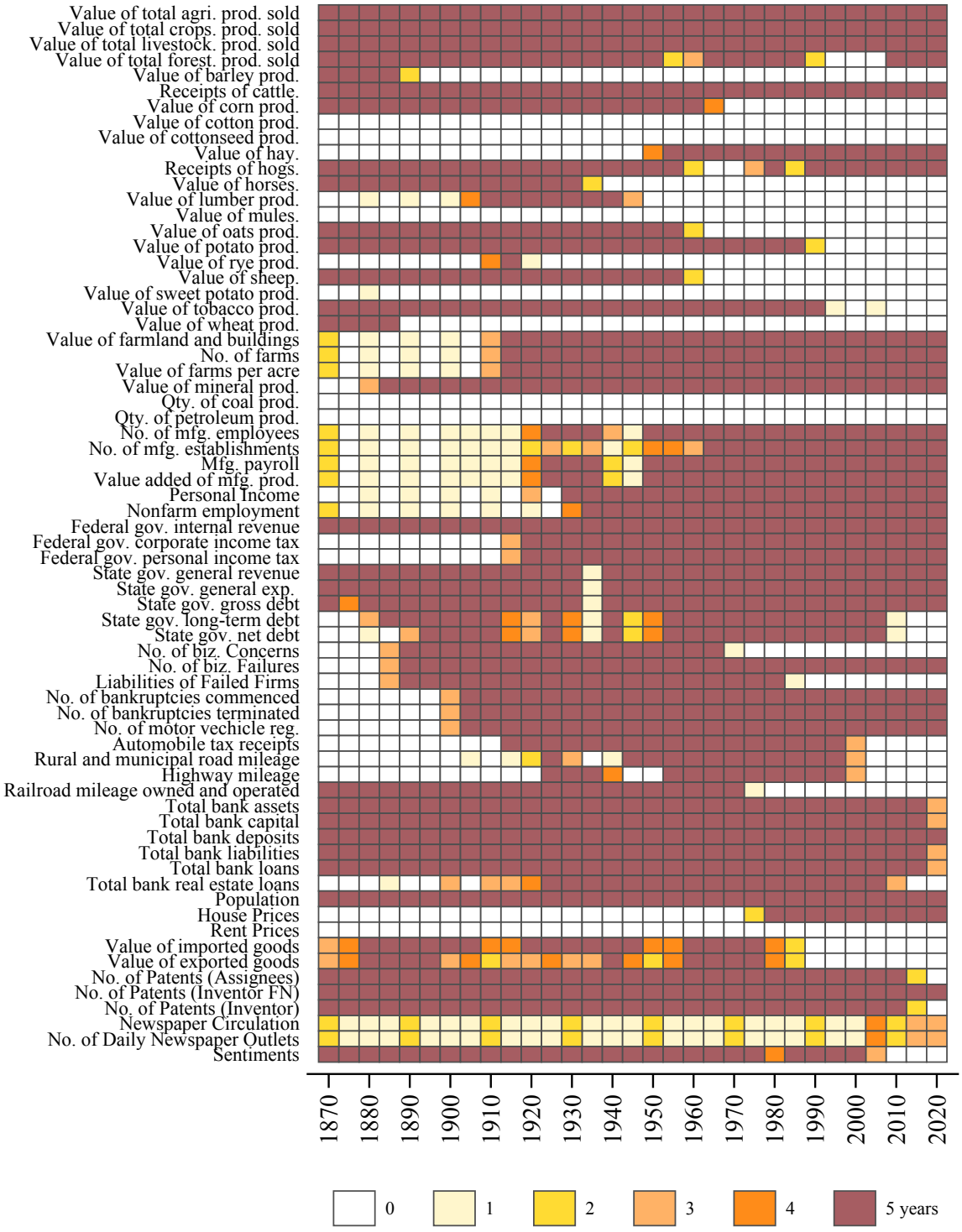
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 8: Availability of Variables – Colorado



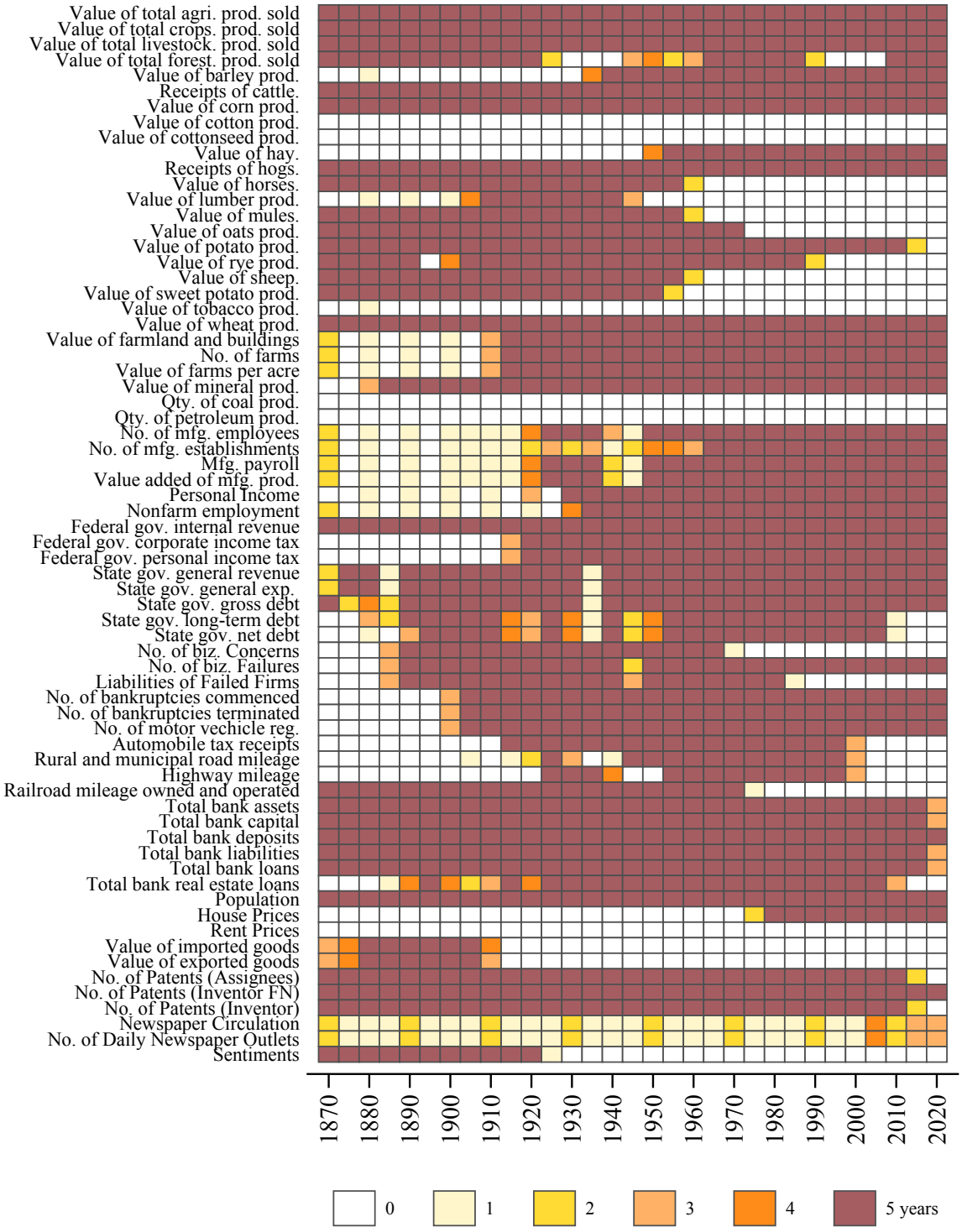
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 9: Availability of Variables – Connecticut



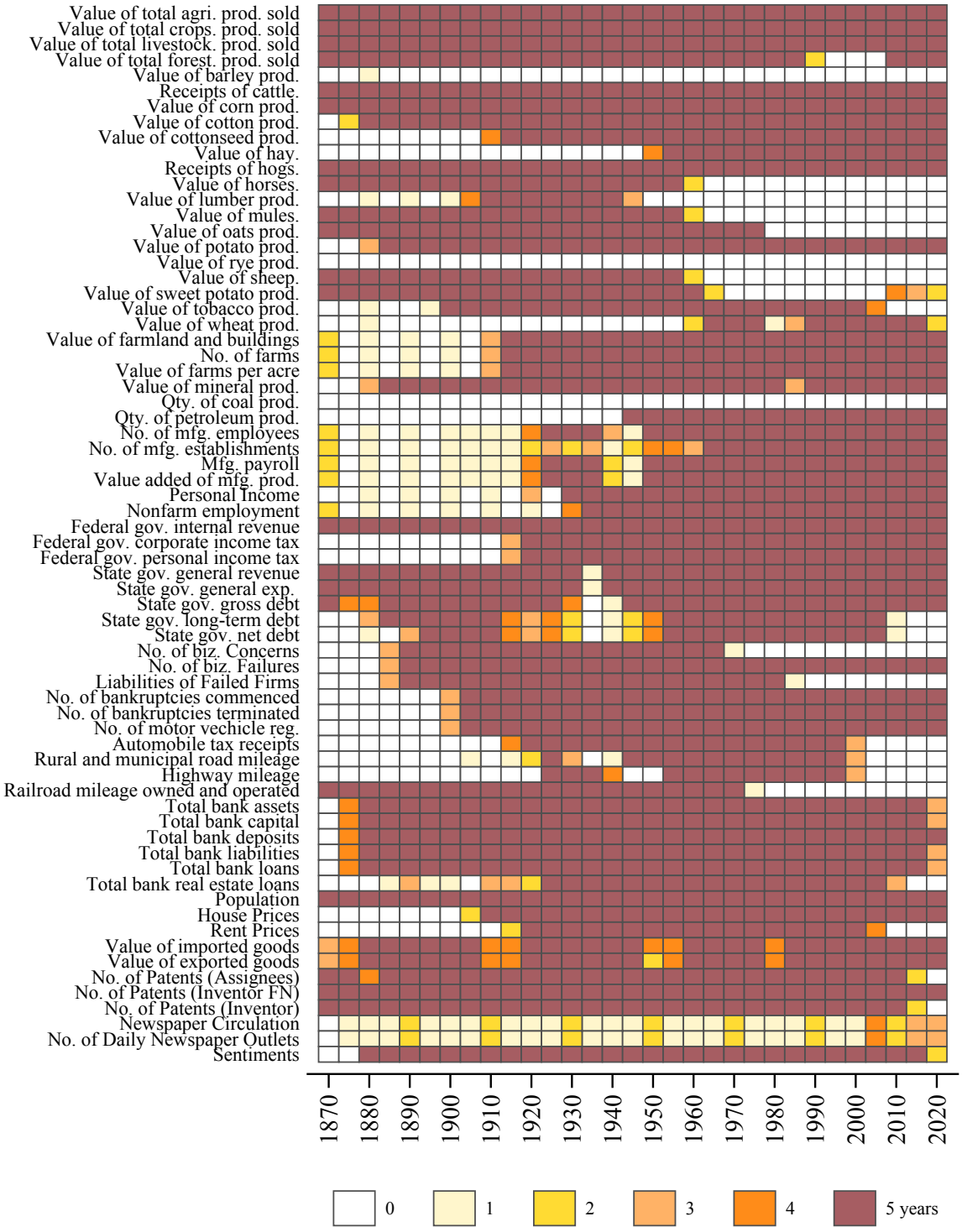
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 10: Availability of Variables – Delaware



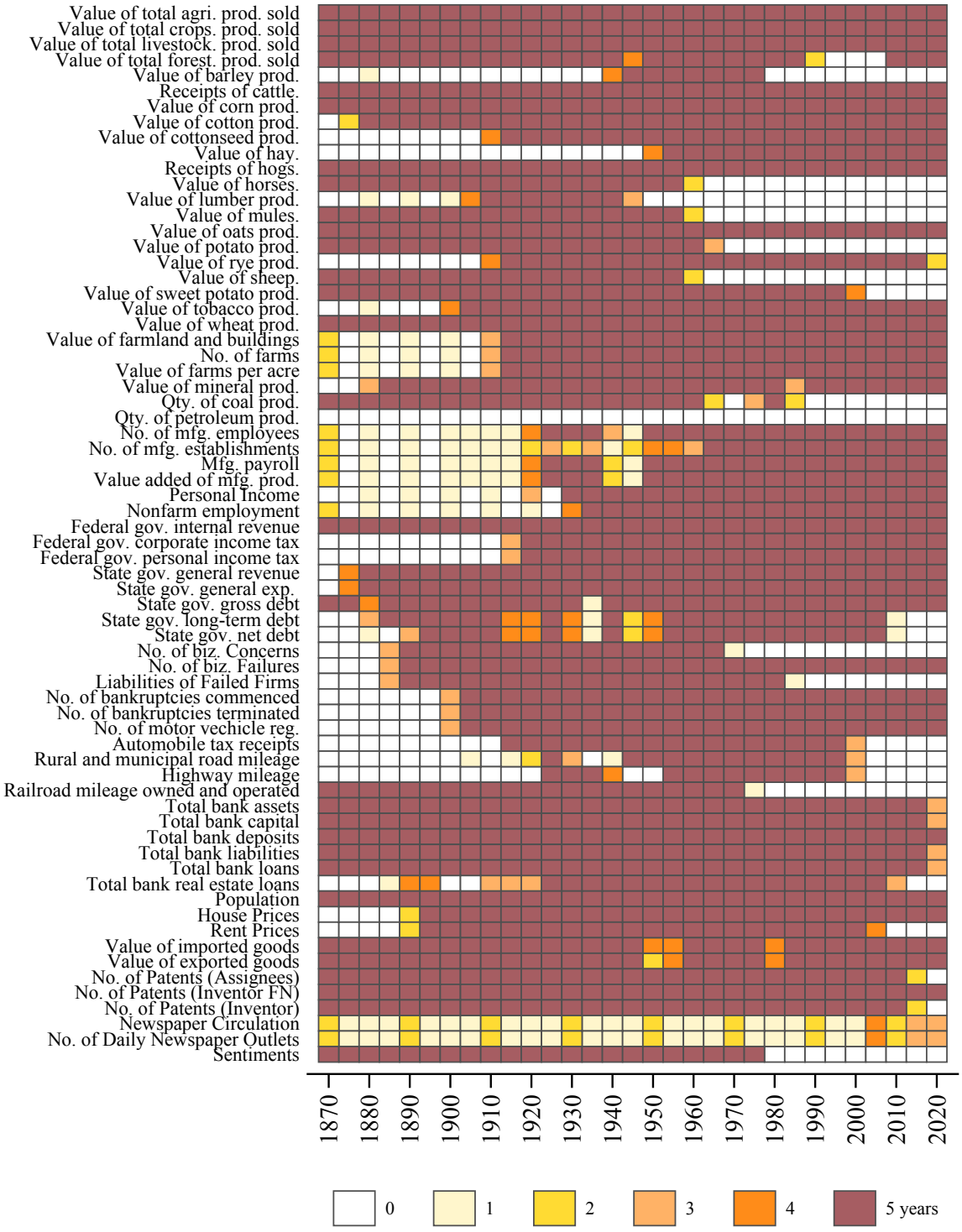
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 11: Availability of Variables – Florida



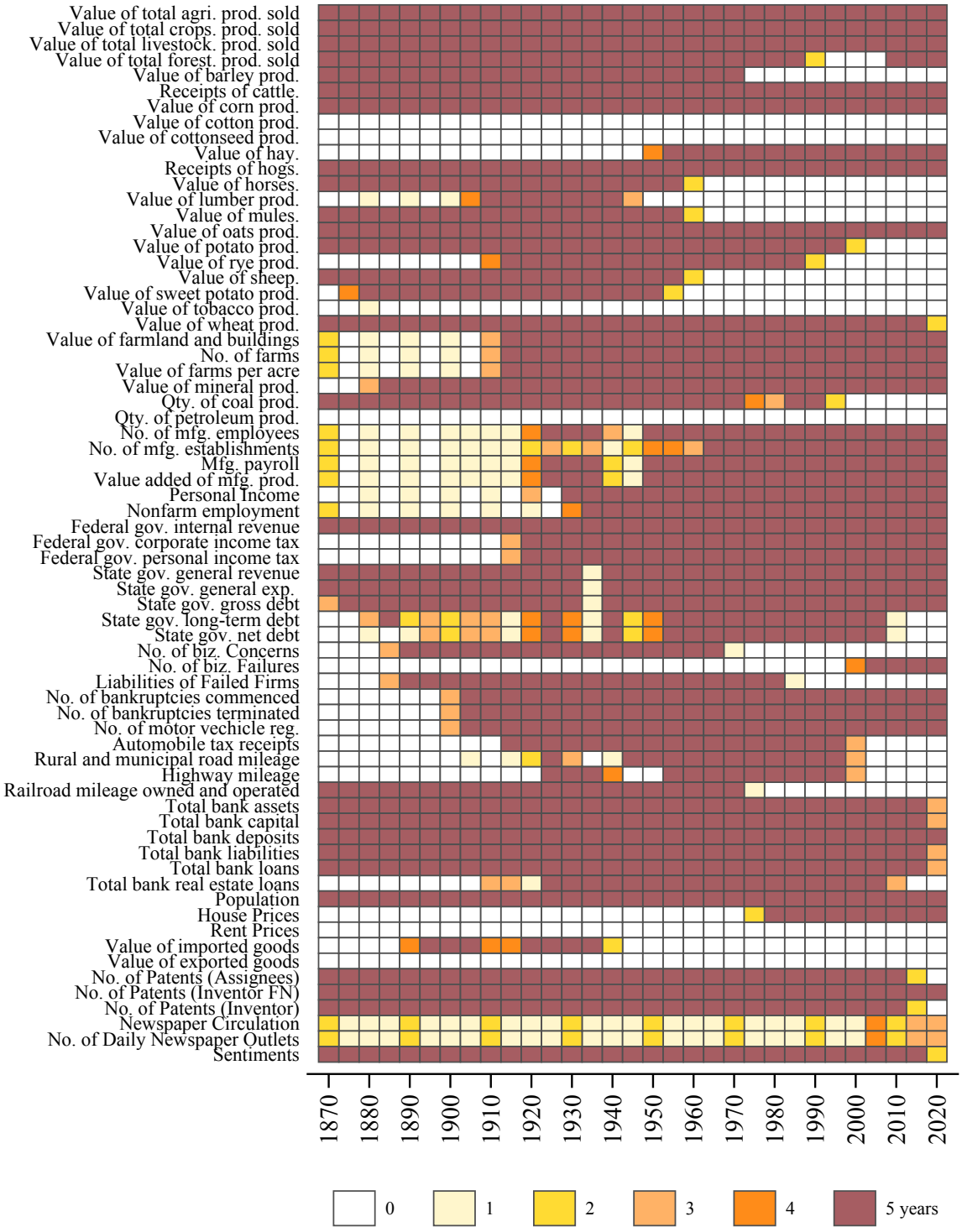
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 12: Availability of Variables – Georgia



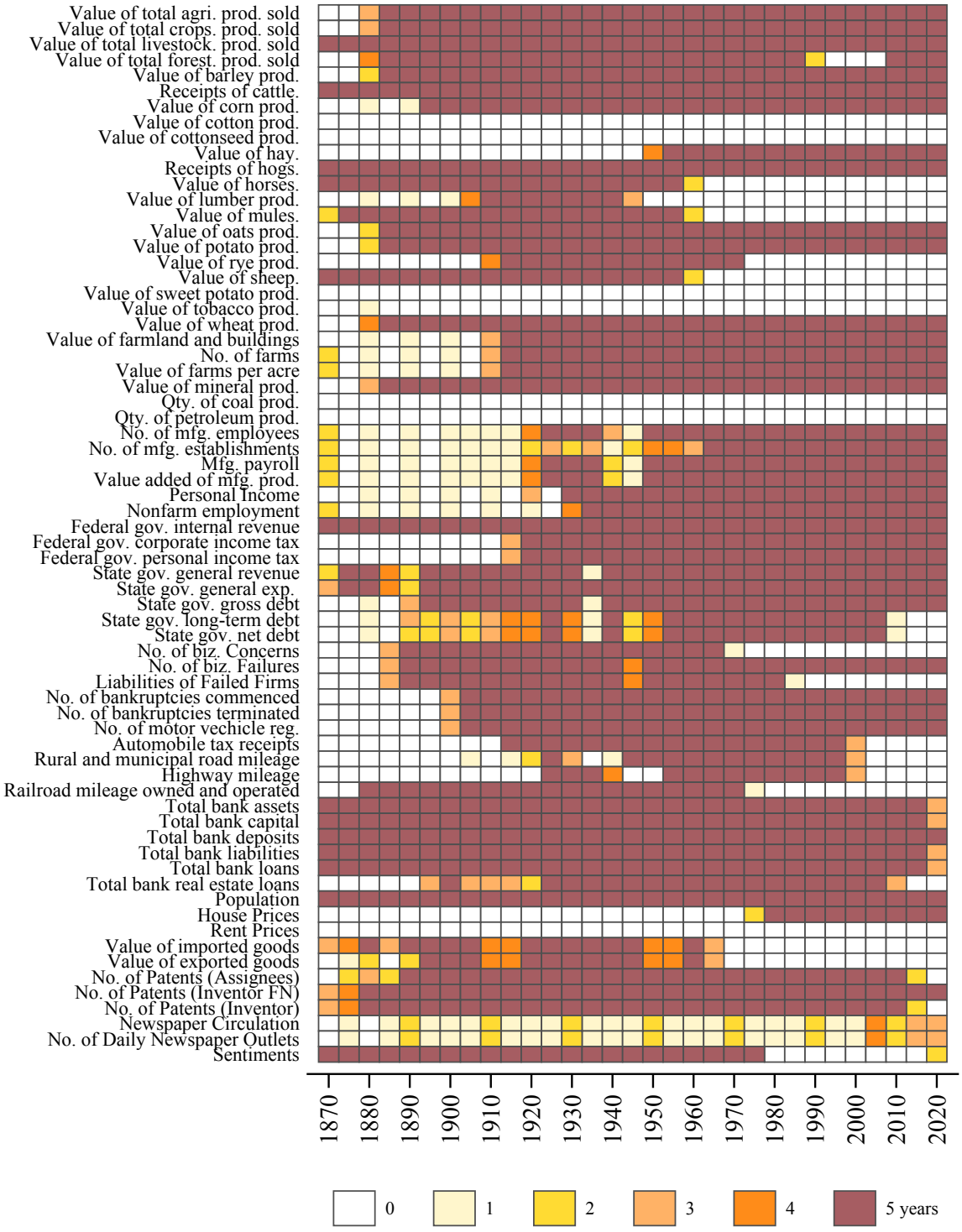
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 13: Availability of Variables – Iowa



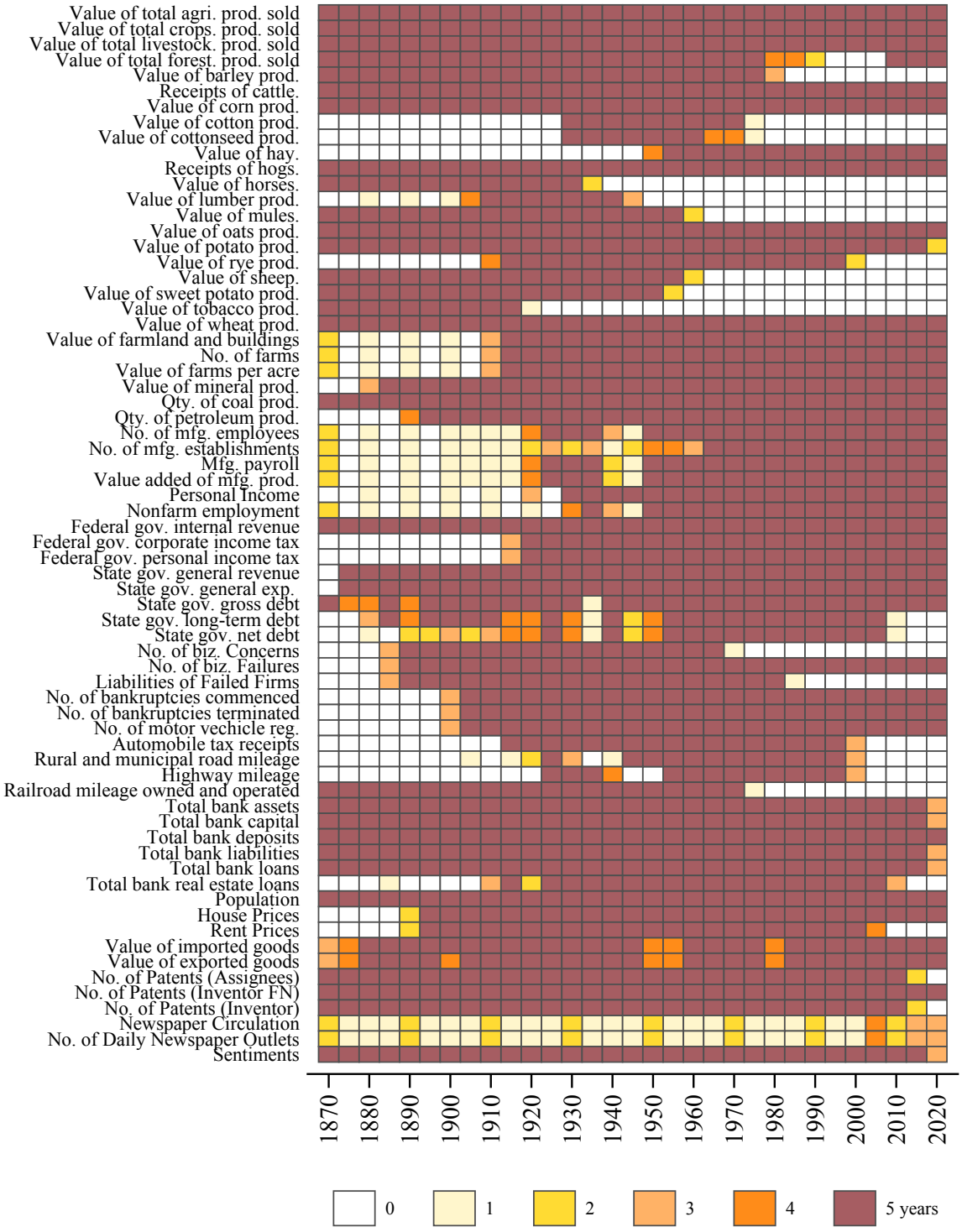
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 14: Availability of Variables – Idaho



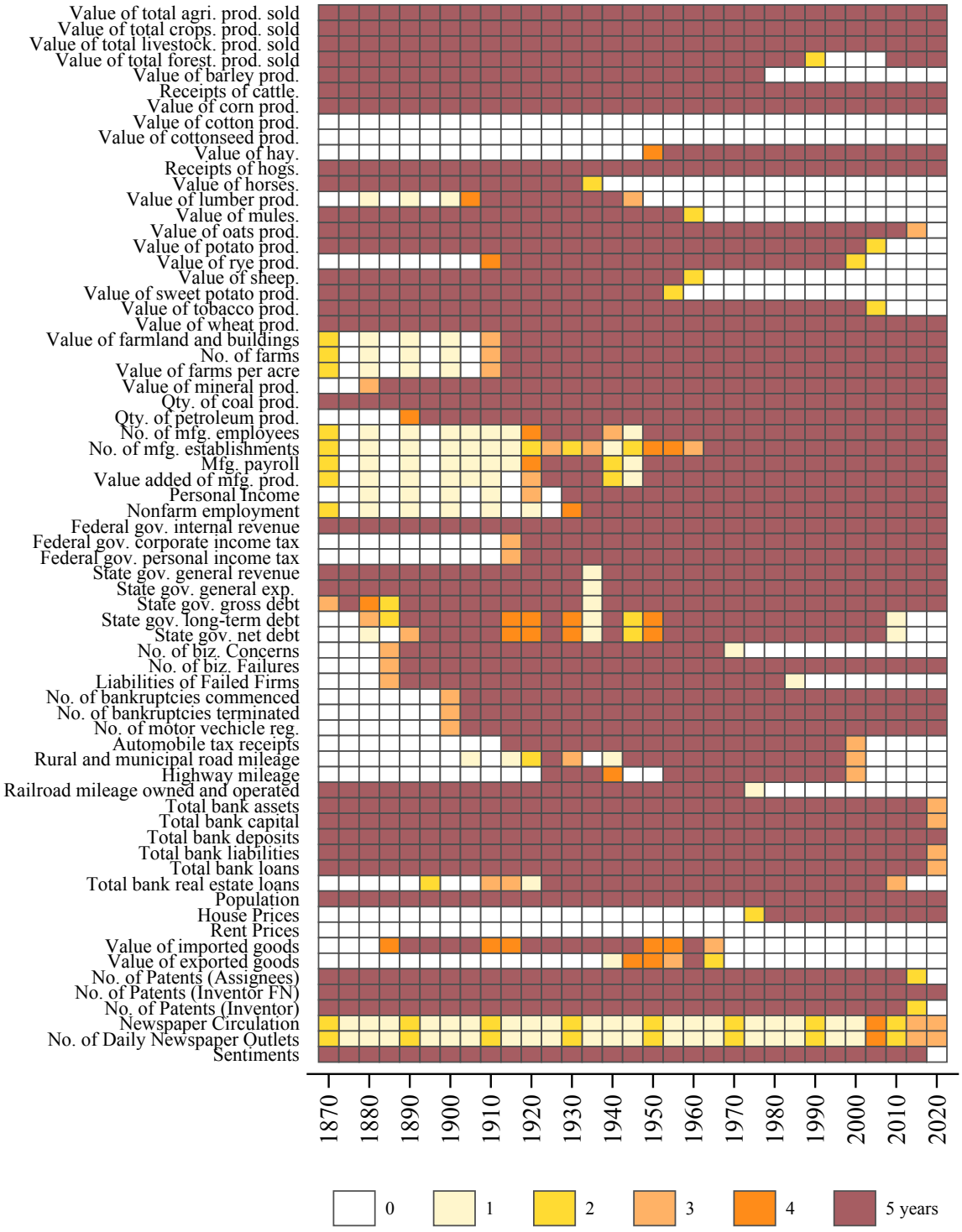
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 15: Availability of Variables – Illinois



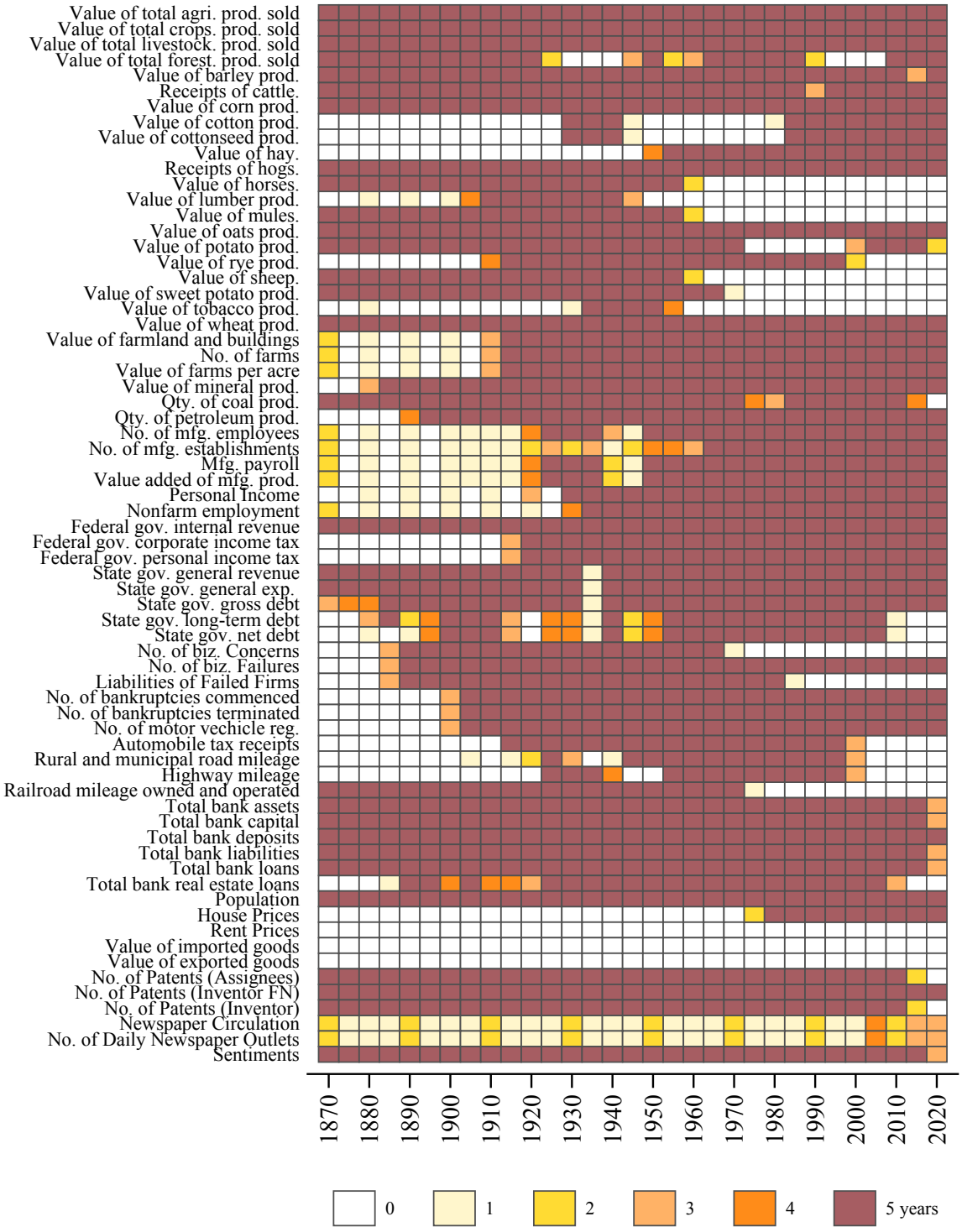
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 16: Availability of Variables – Indiana



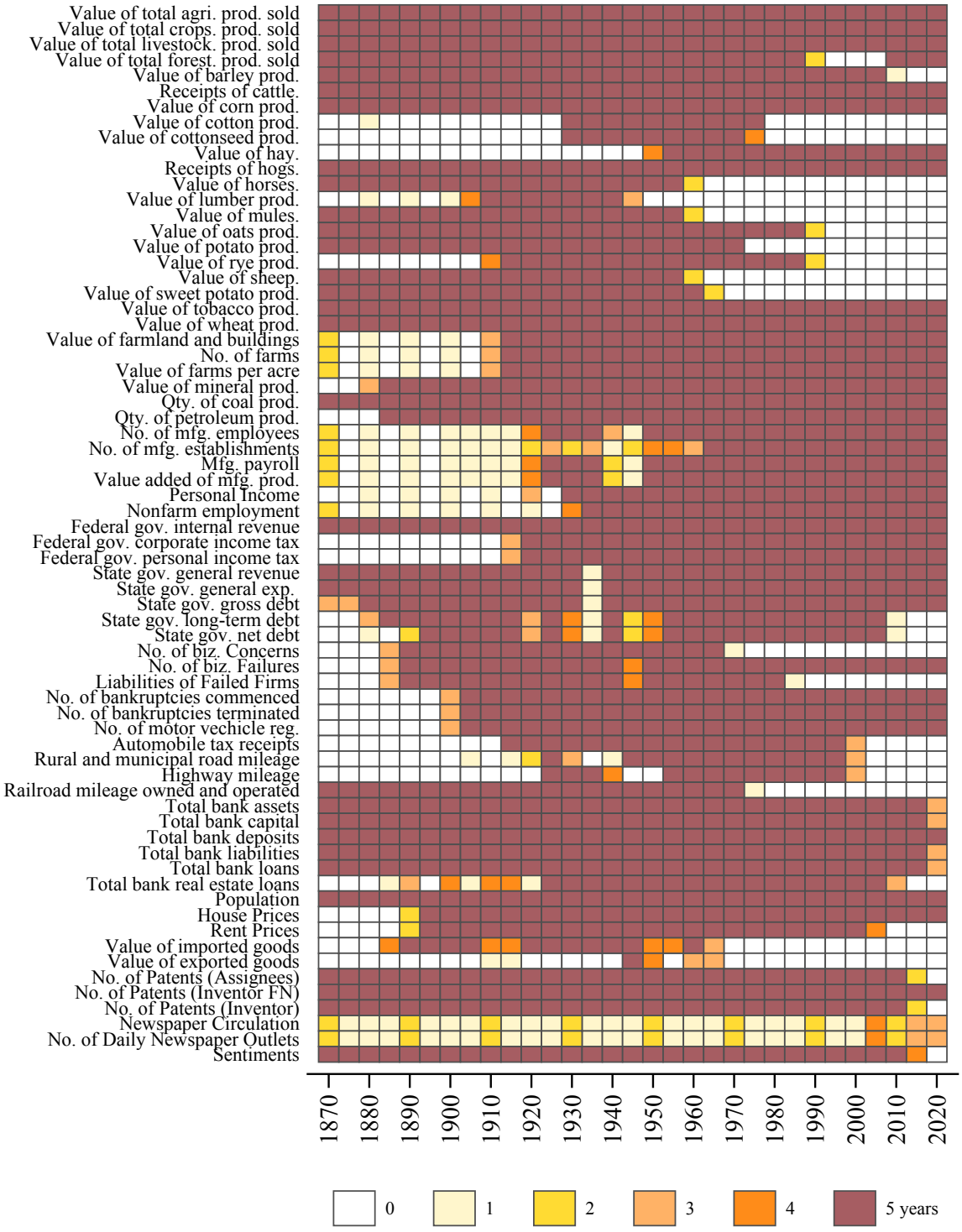
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 17: Availability of Variables – Kansas



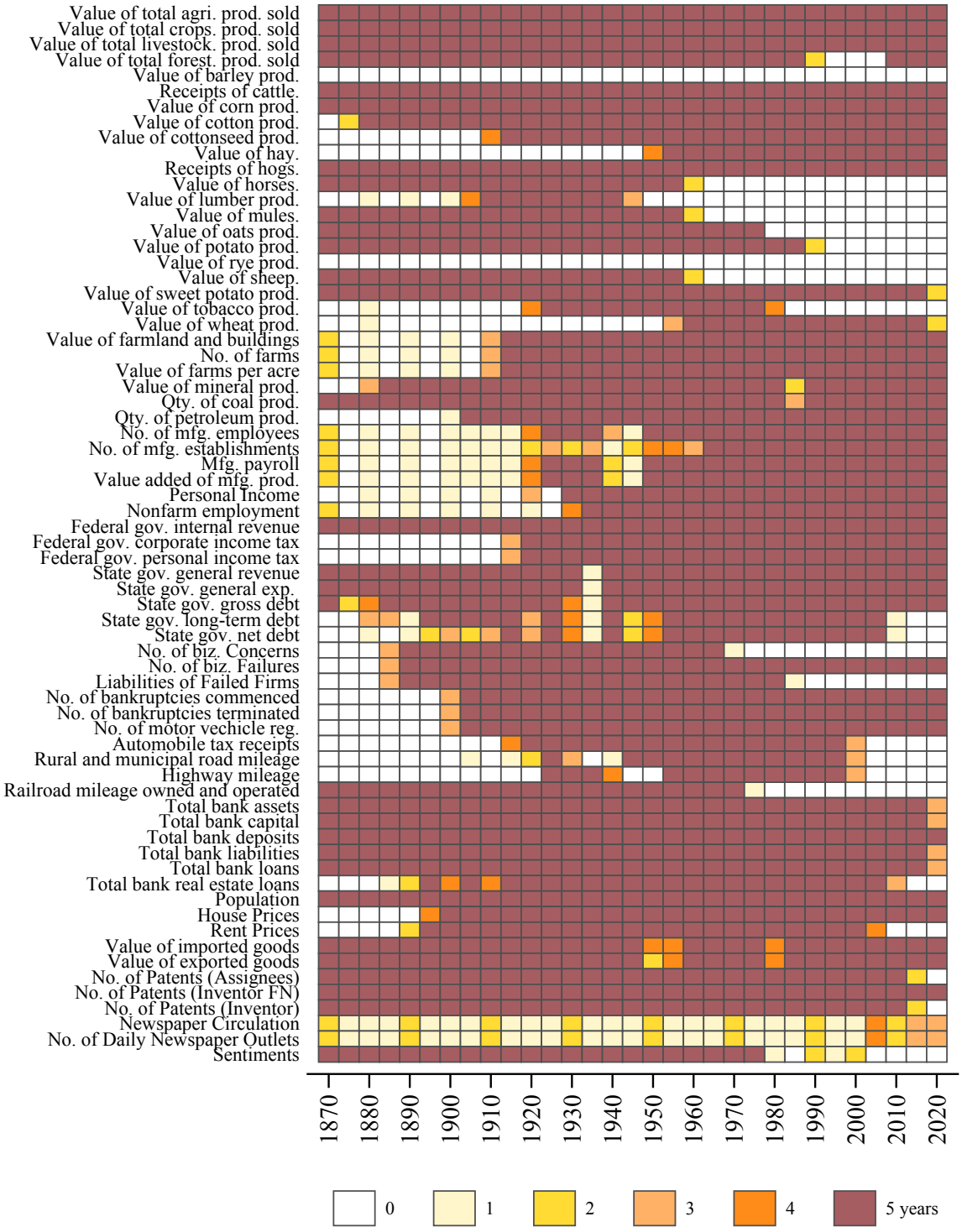
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 18: Availability of Variables – Kentucky



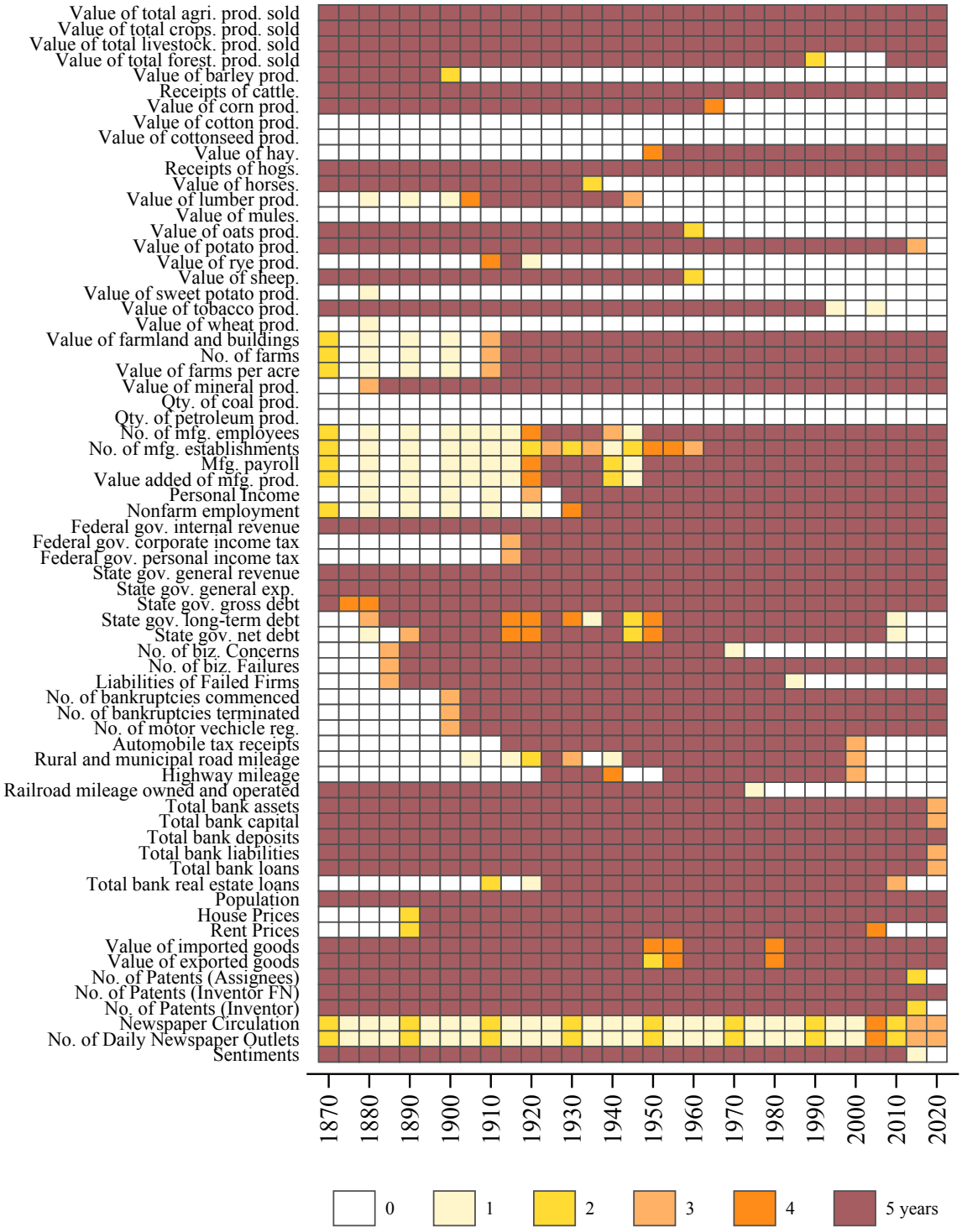
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 19: Availability of Variables – Louisiana



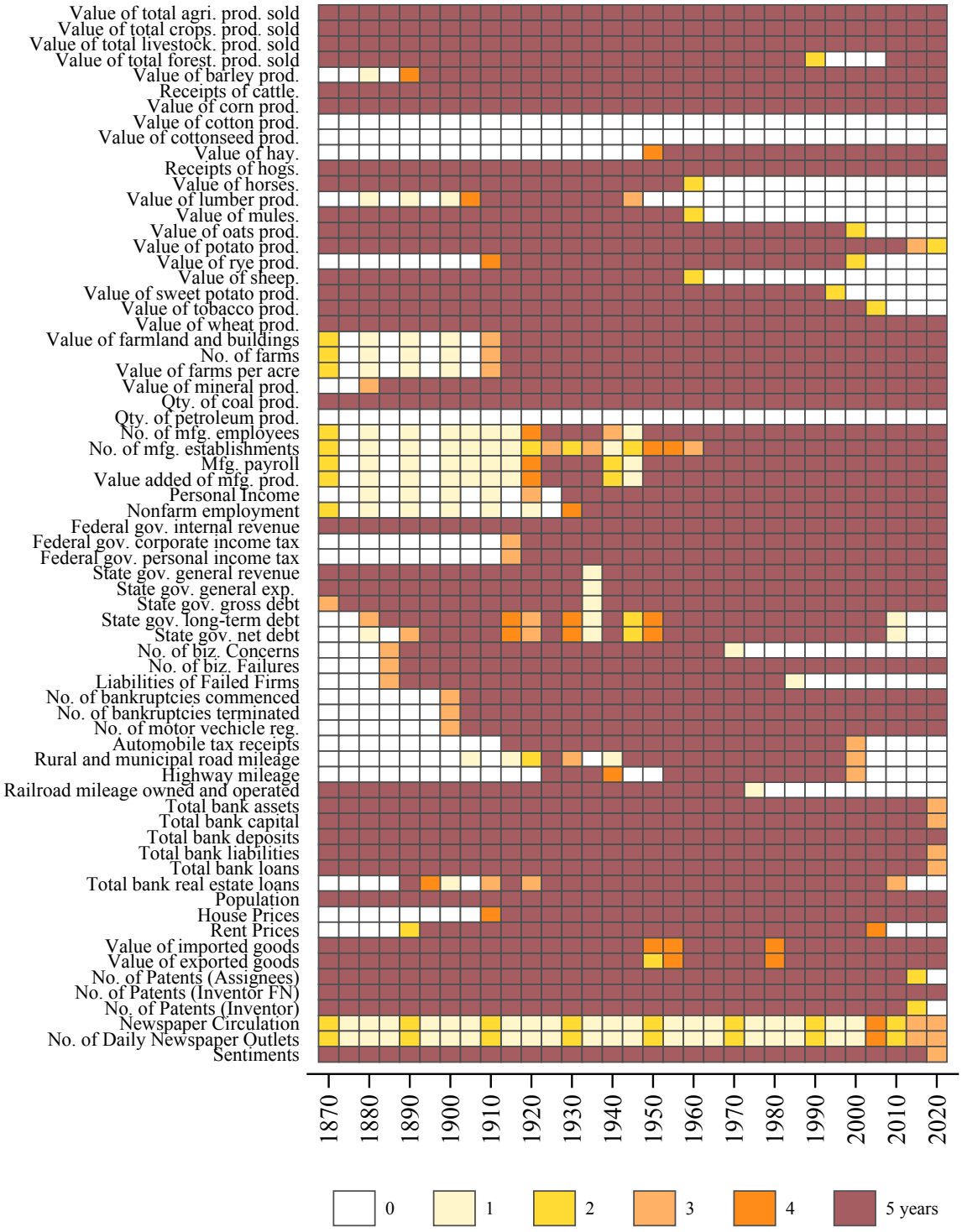
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 20: Availability of Variables – Massachusetts



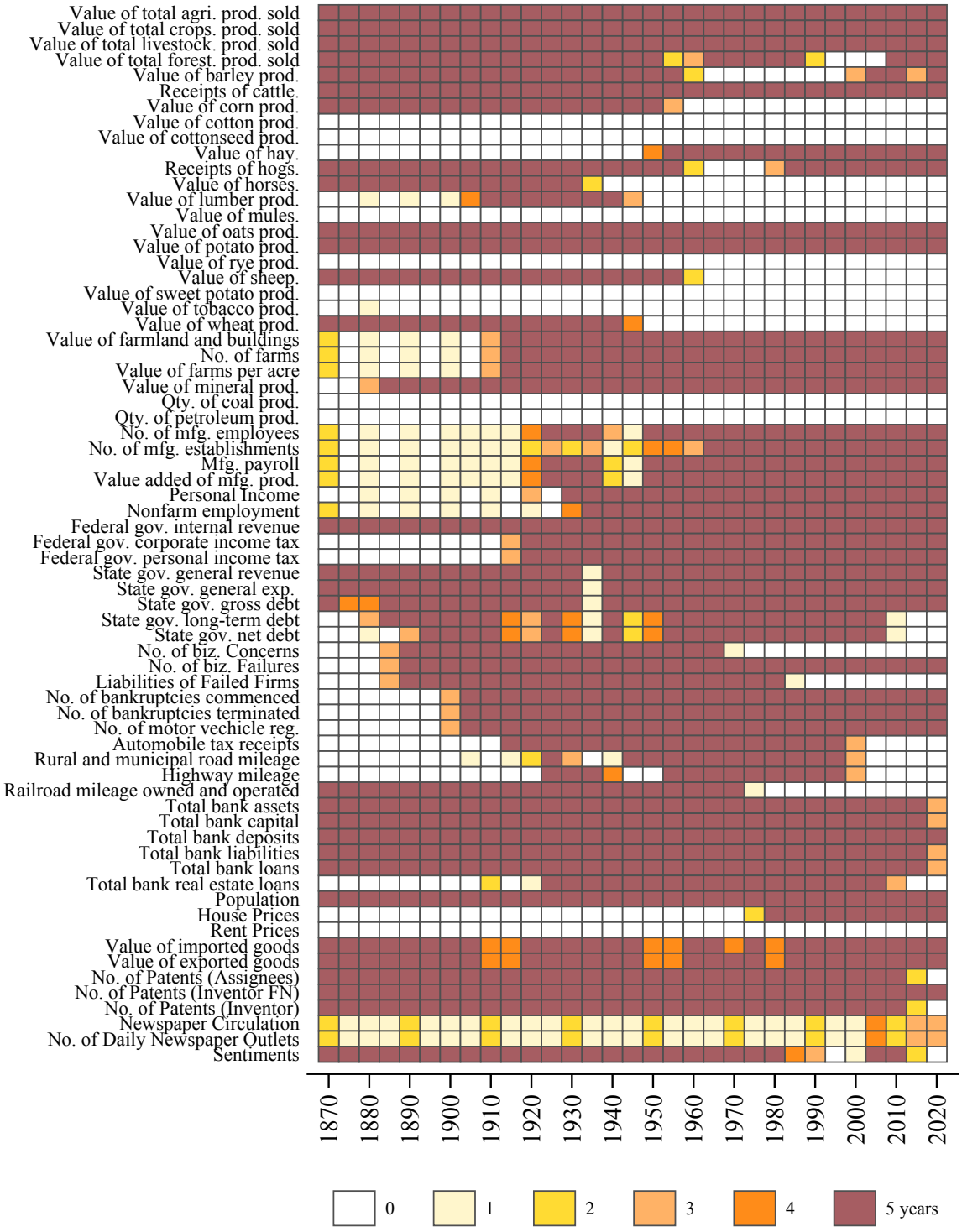
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 21: Availability of Variables – Maryland



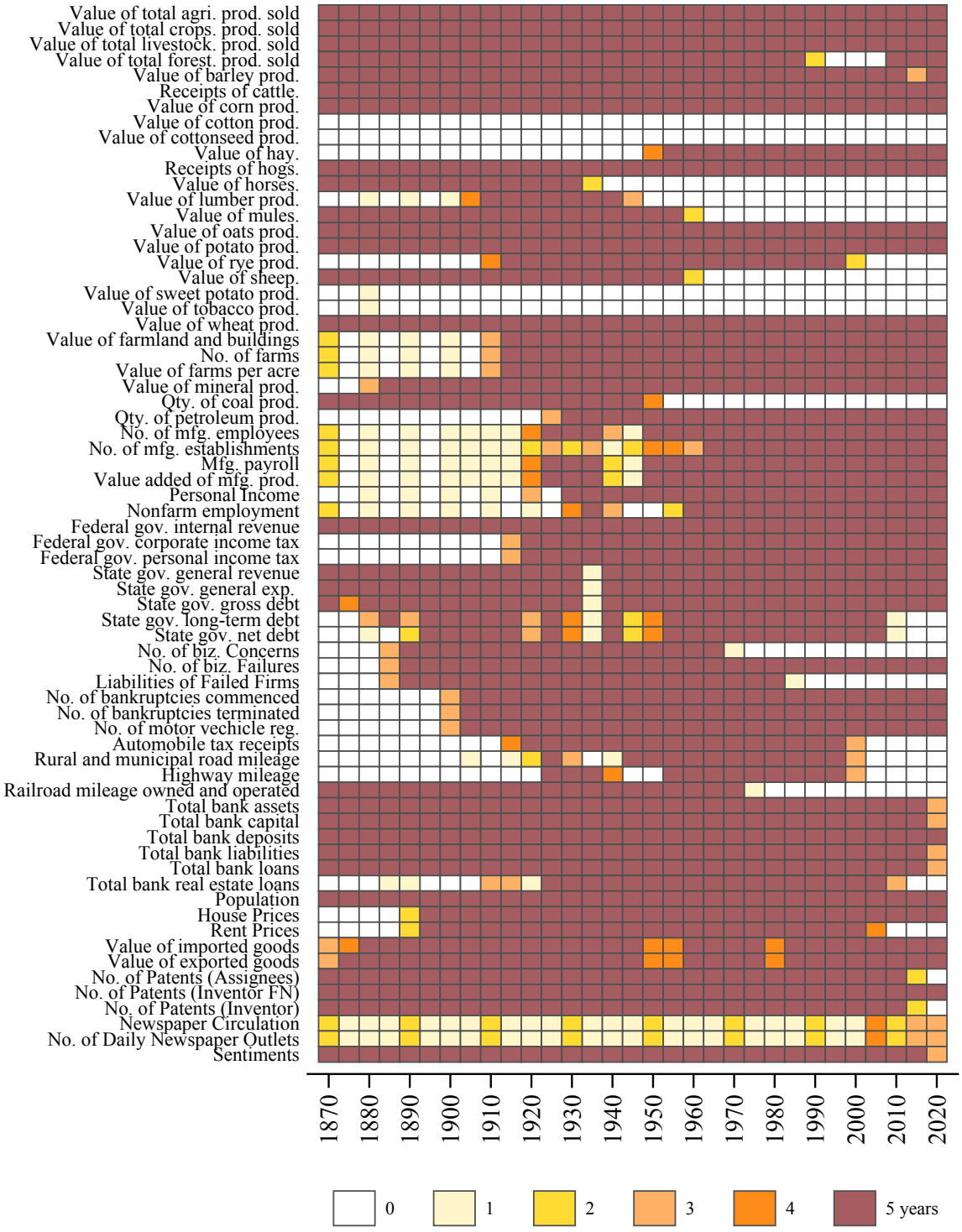
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 22: Availability of Variables – Maine



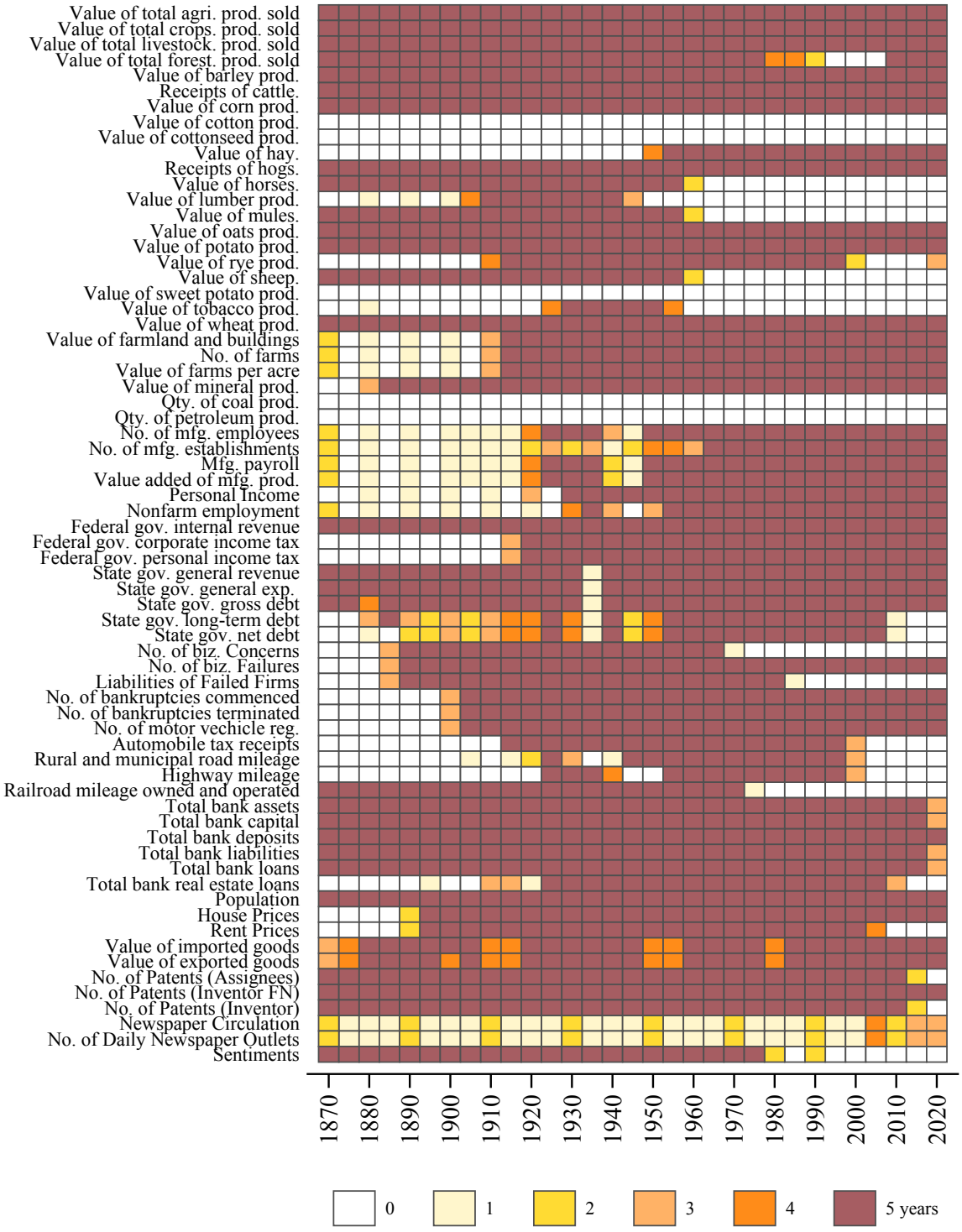
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 23: Availability of Variables – Michigan



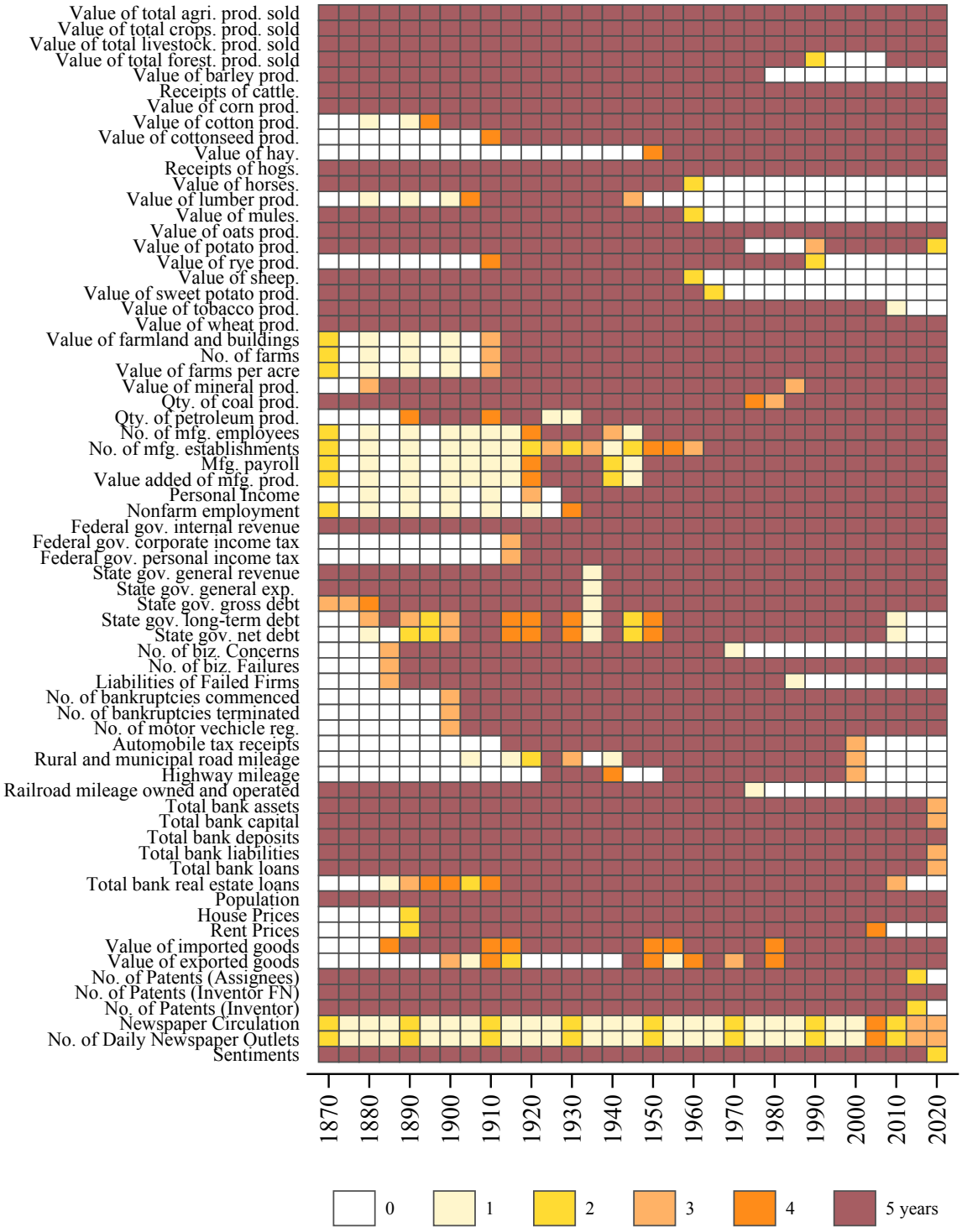
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 24: Availability of Variables – Minnesota



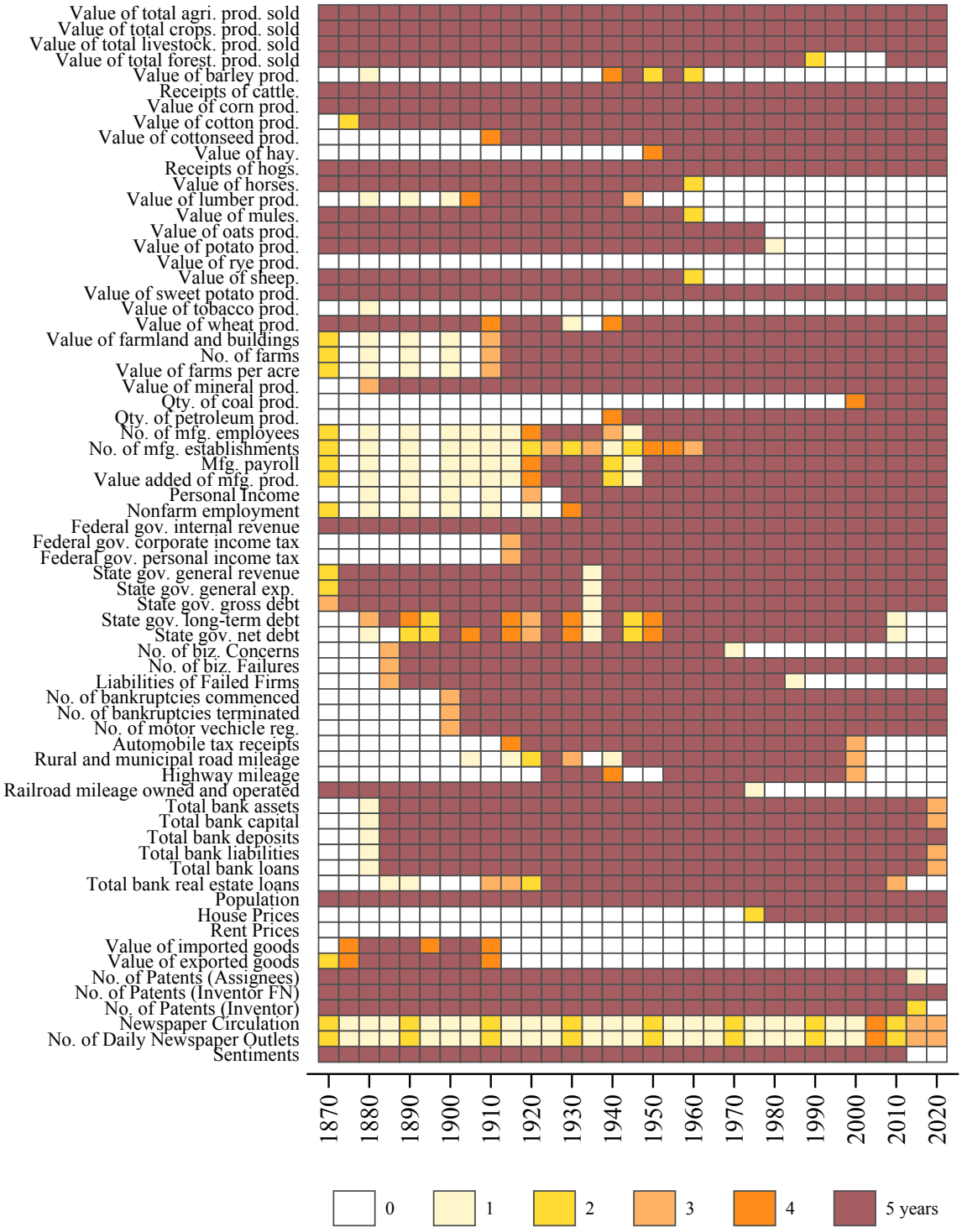
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 25: Availability of Variables – Missouri



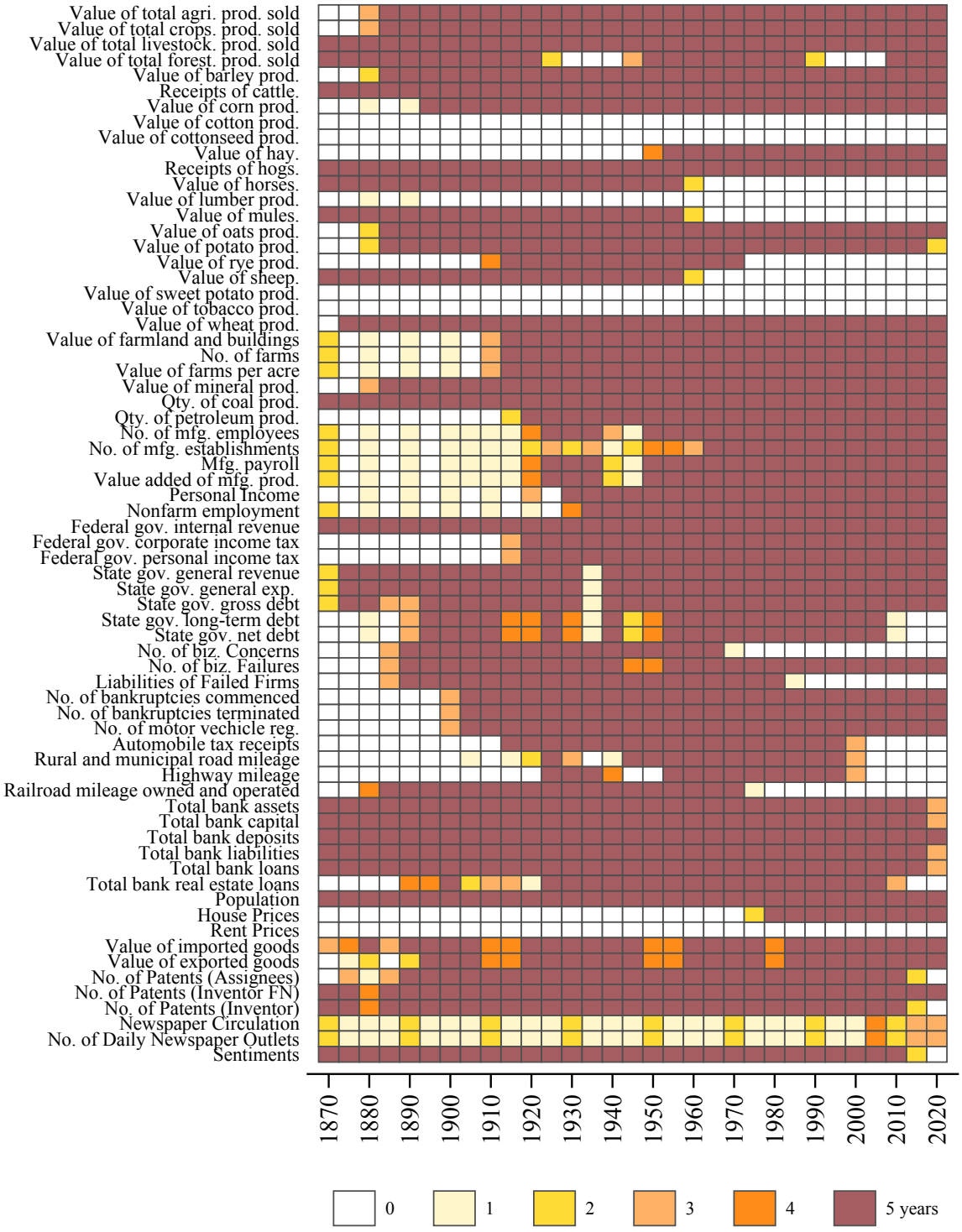
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 26: Availability of Variables – Mississippi



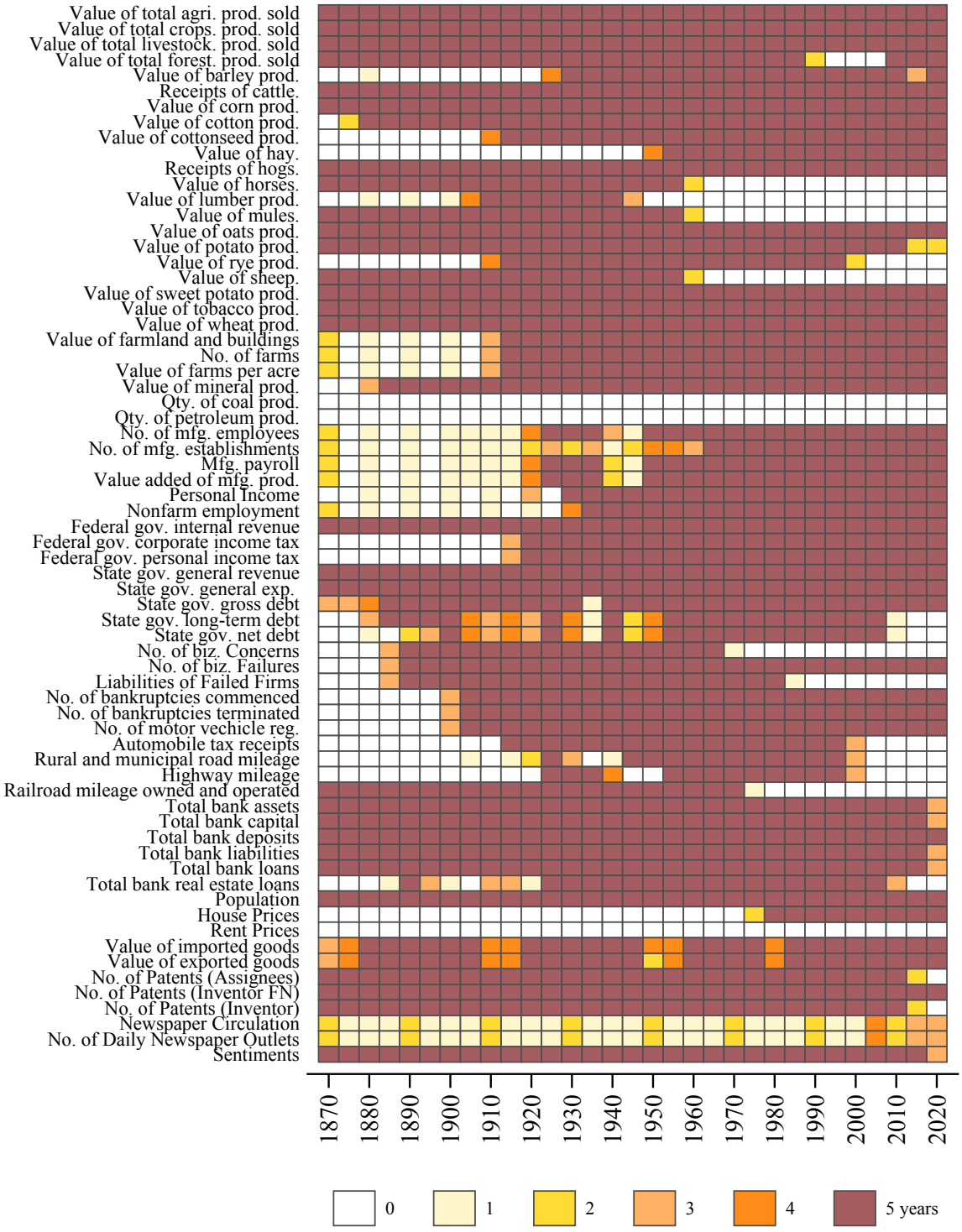
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 27: Availability of Variables – Montana



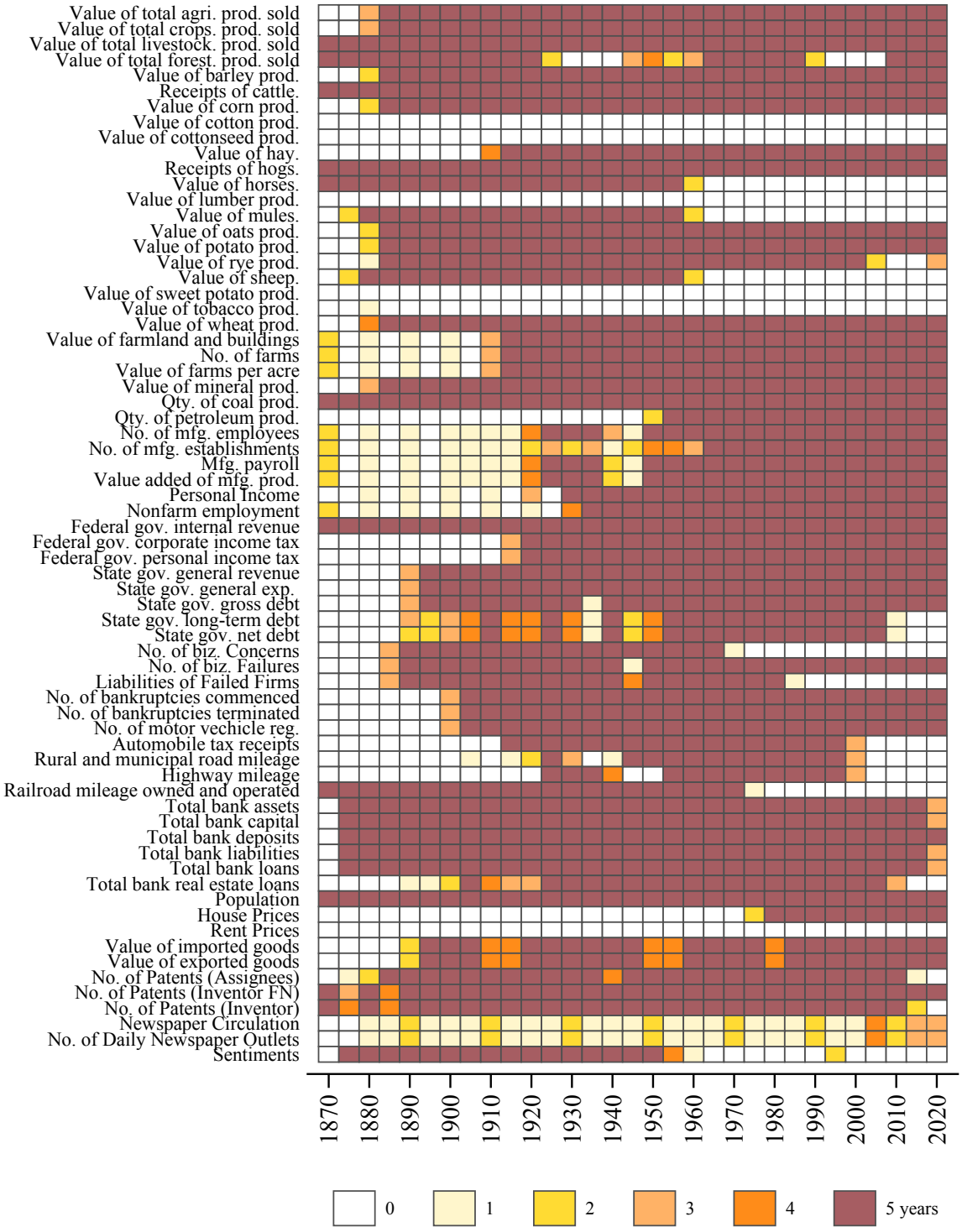
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 28: Availability of Variables – North Carolina



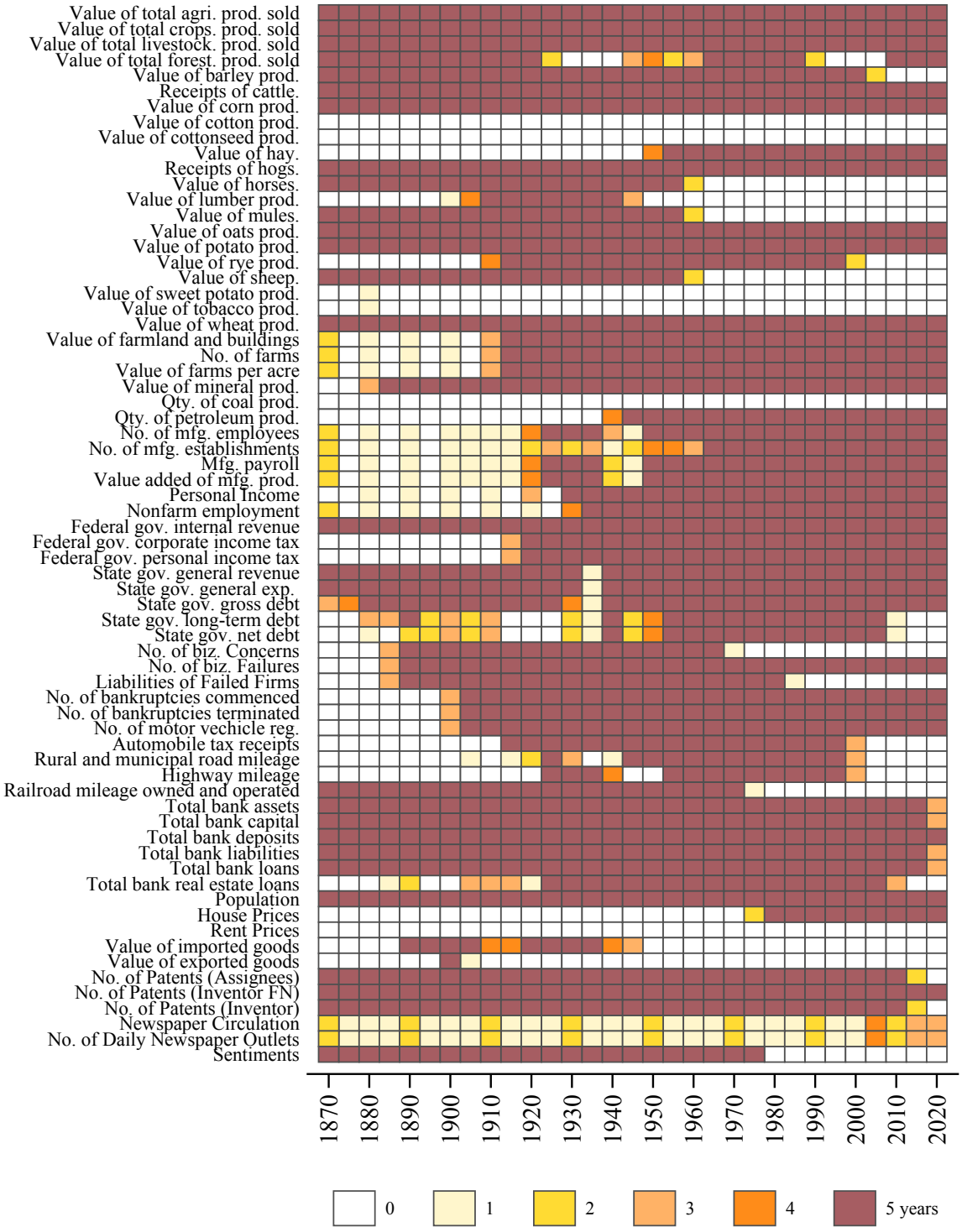
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 29: Availability of Variables – North Dakota



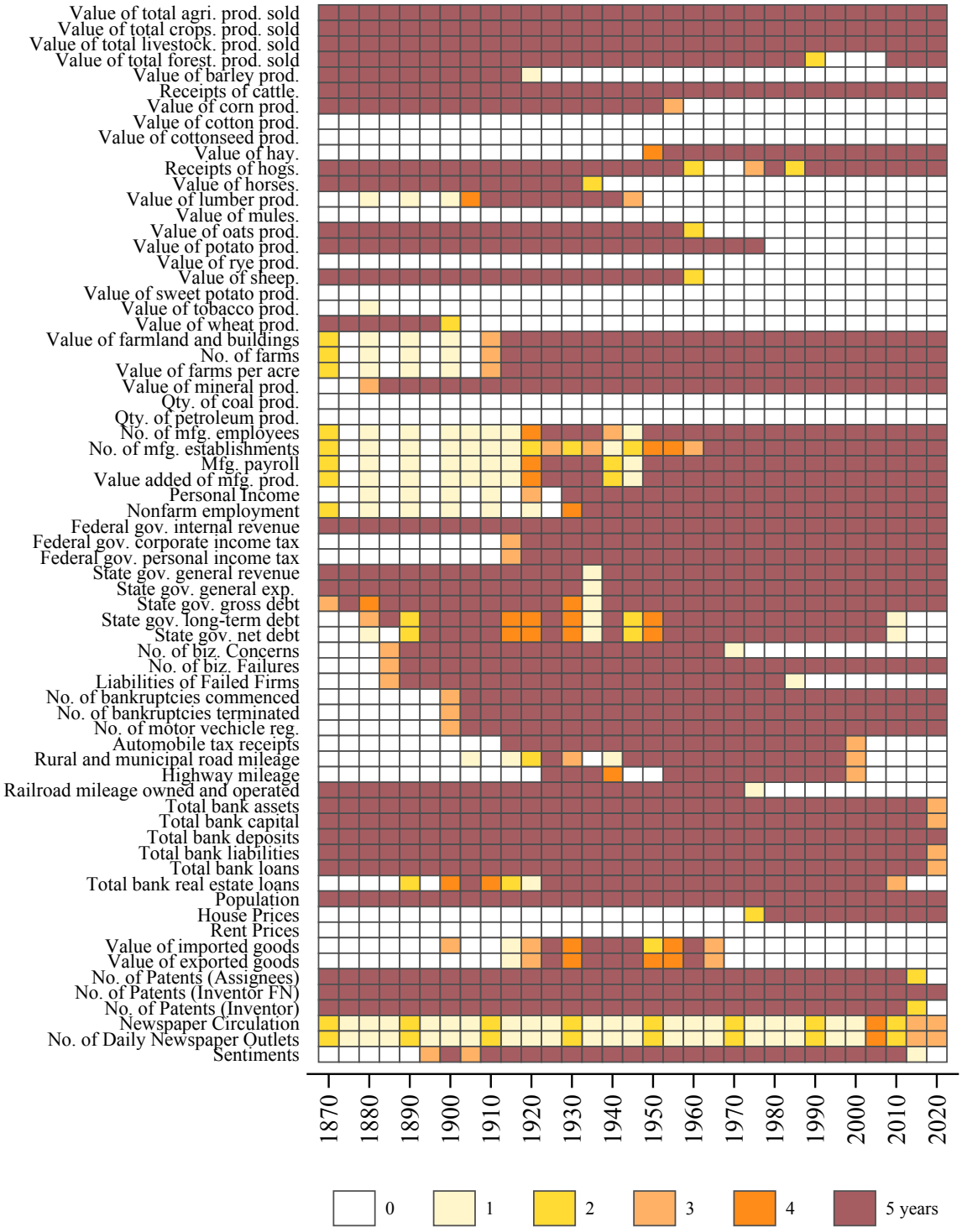
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 30: Availability of Variables – Nebraska



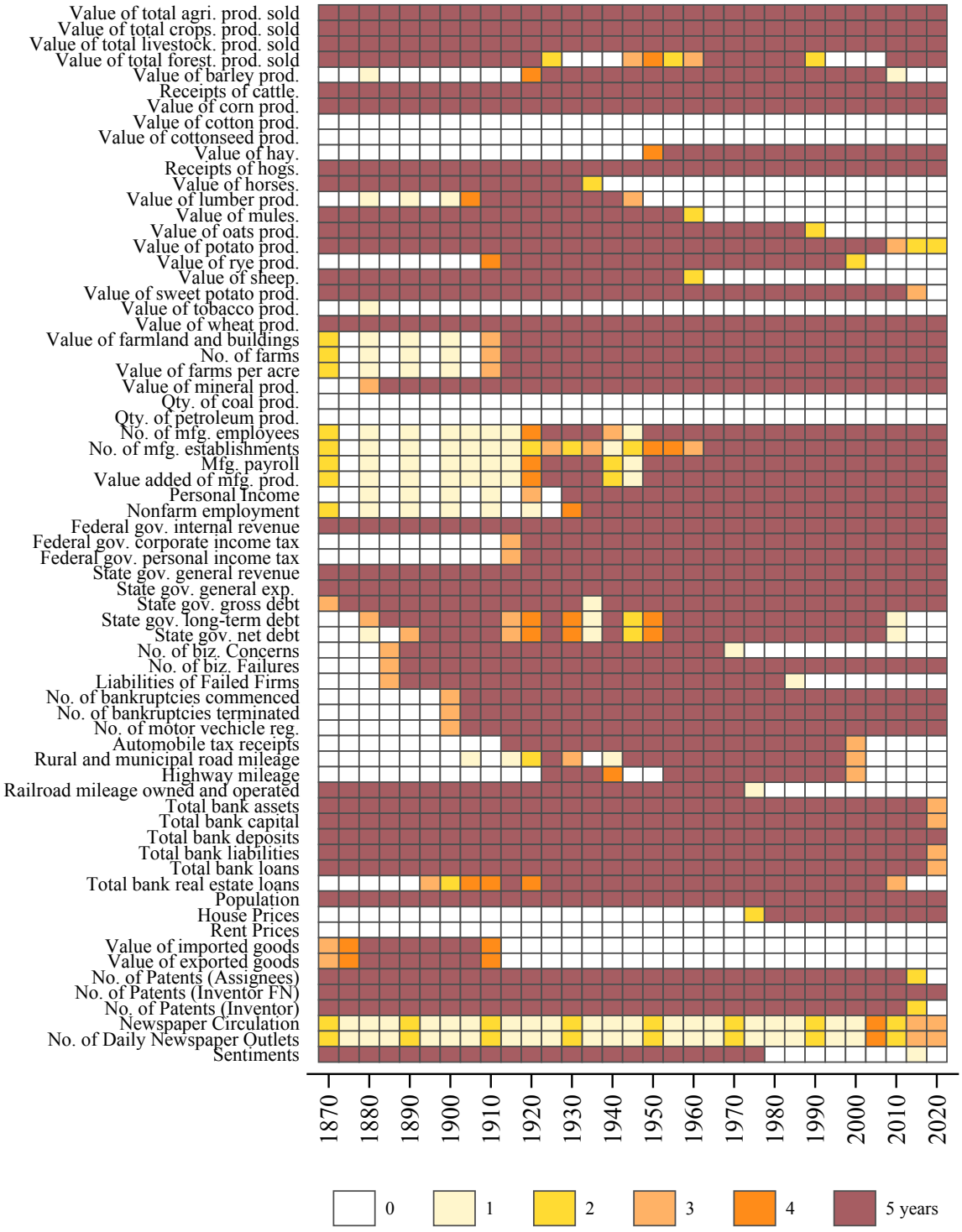
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 31: Availability of Variables – New Hampshire



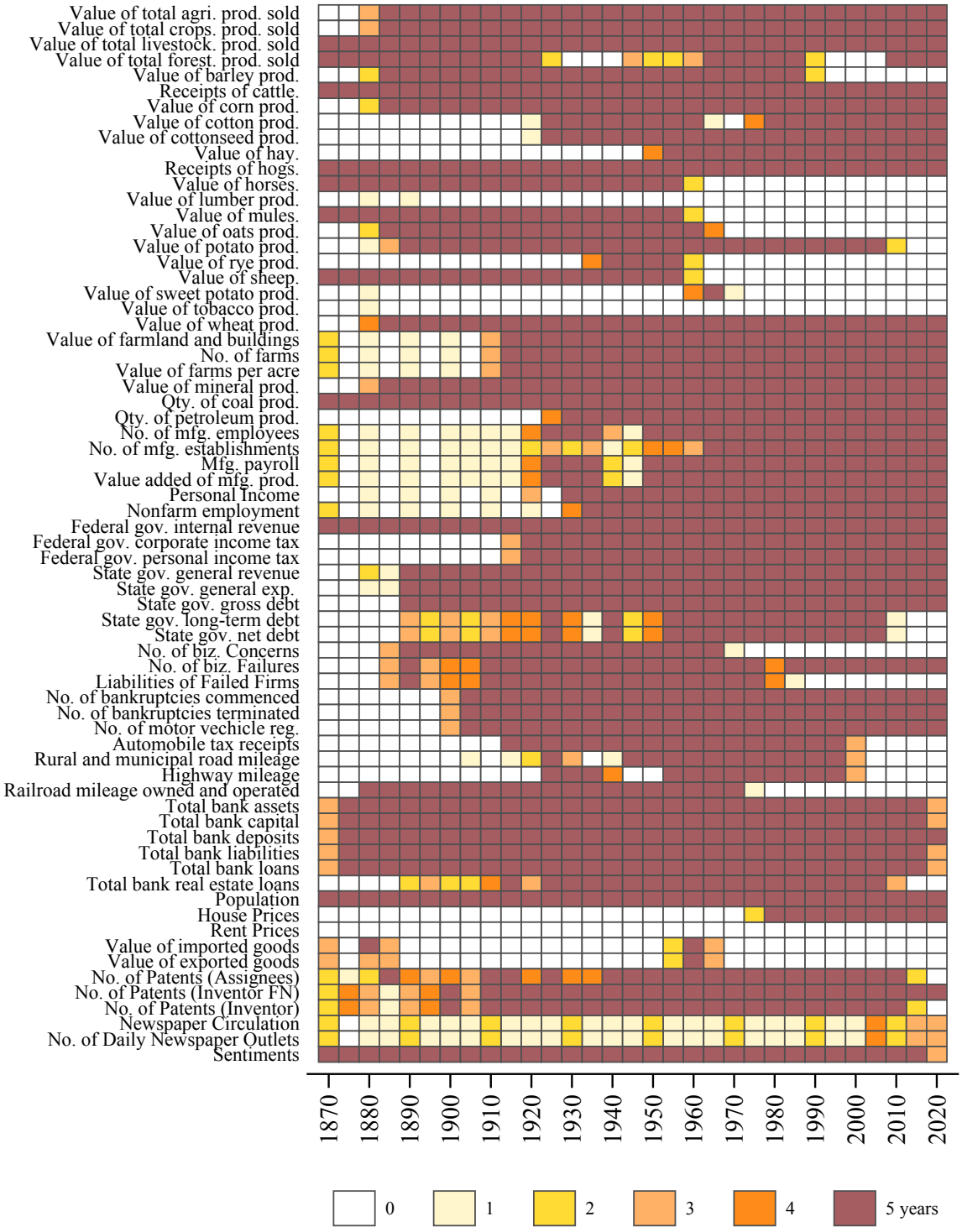
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 32: Availability of Variables – New Jersey



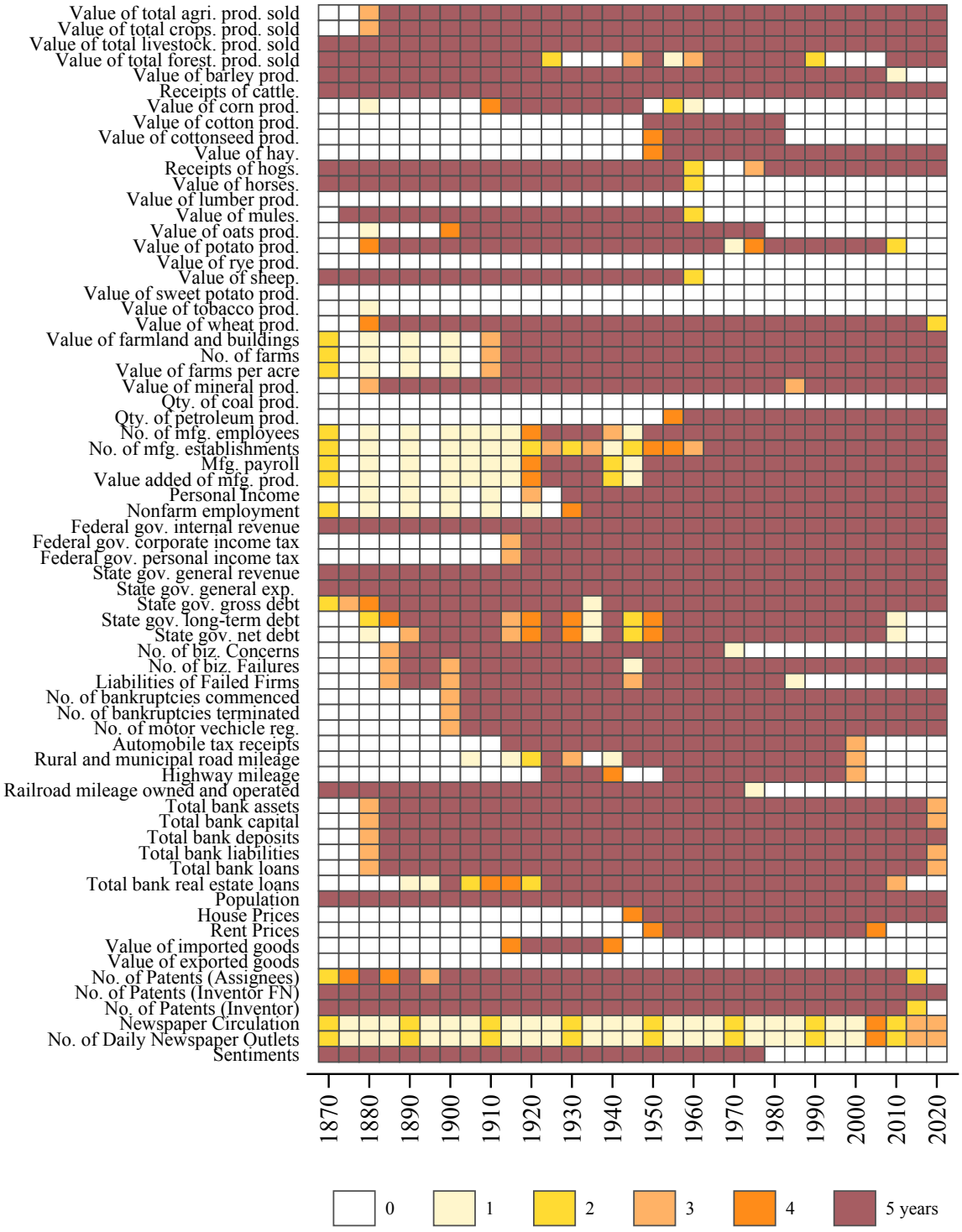
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 33: Availability of Variables – New Mexico



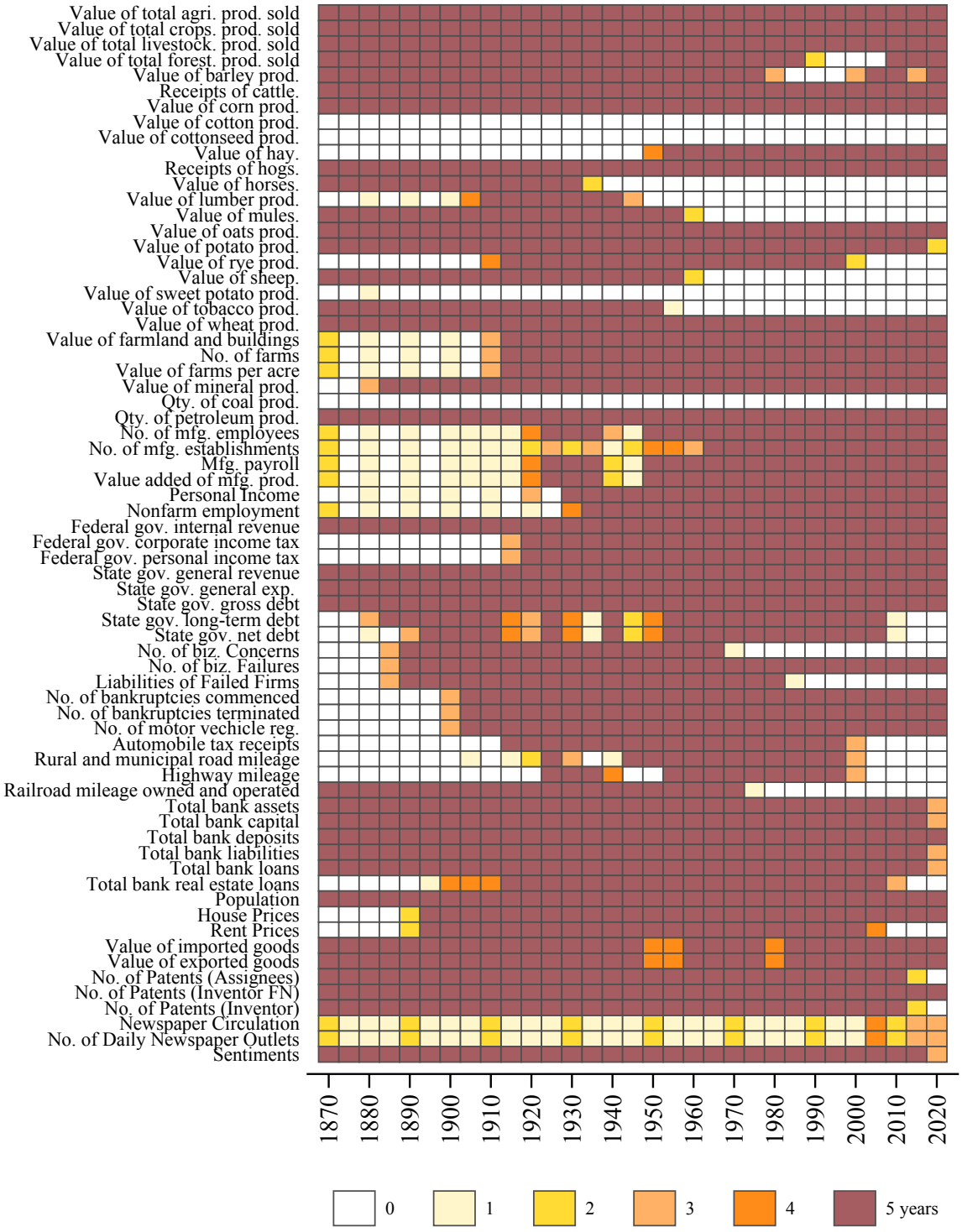
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 34: Availability of Variables – Nevada



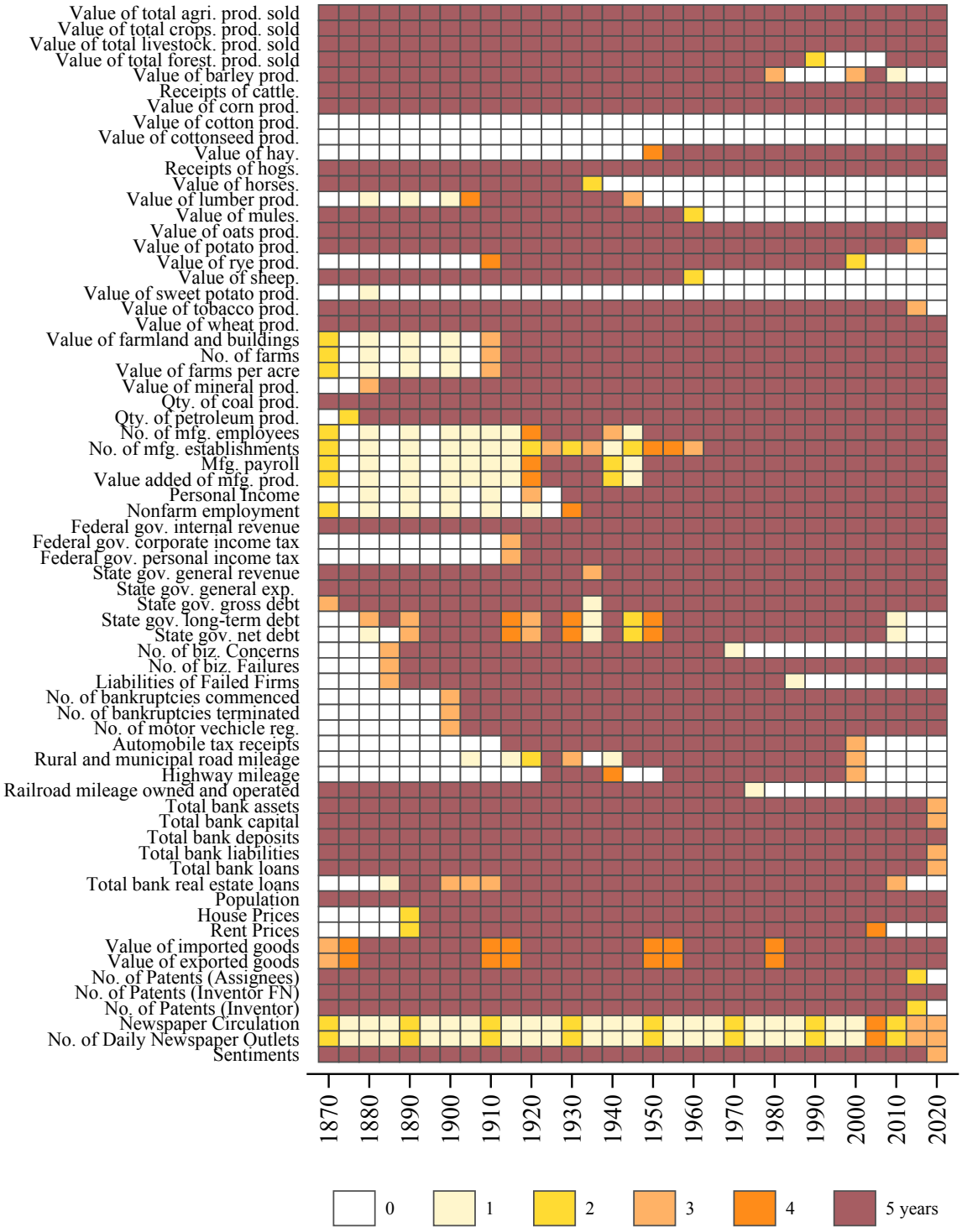
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 35: Availability of Variables – New York



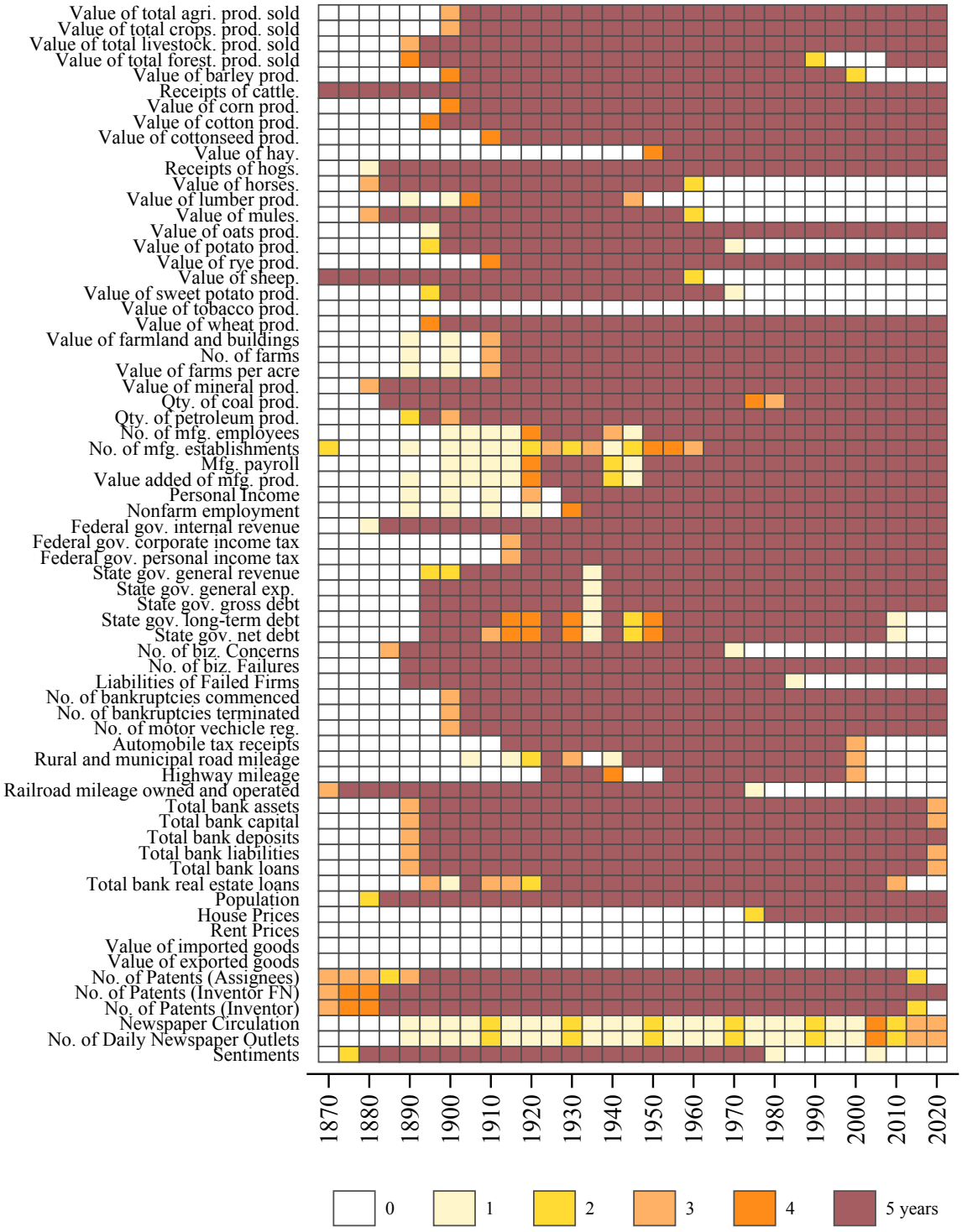
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 36: Availability of Variables – Ohio



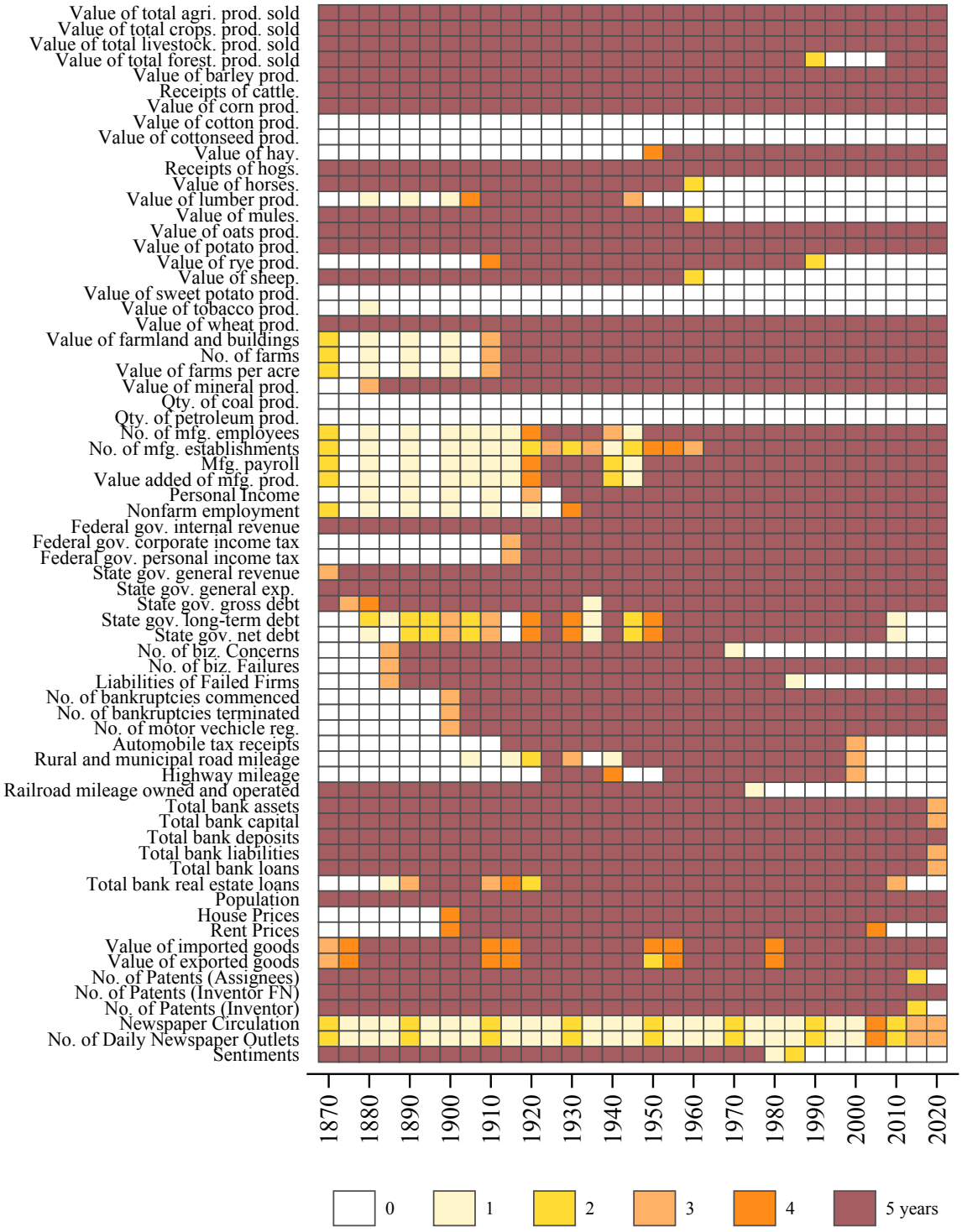
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 37: Availability of Variables – Oklahoma



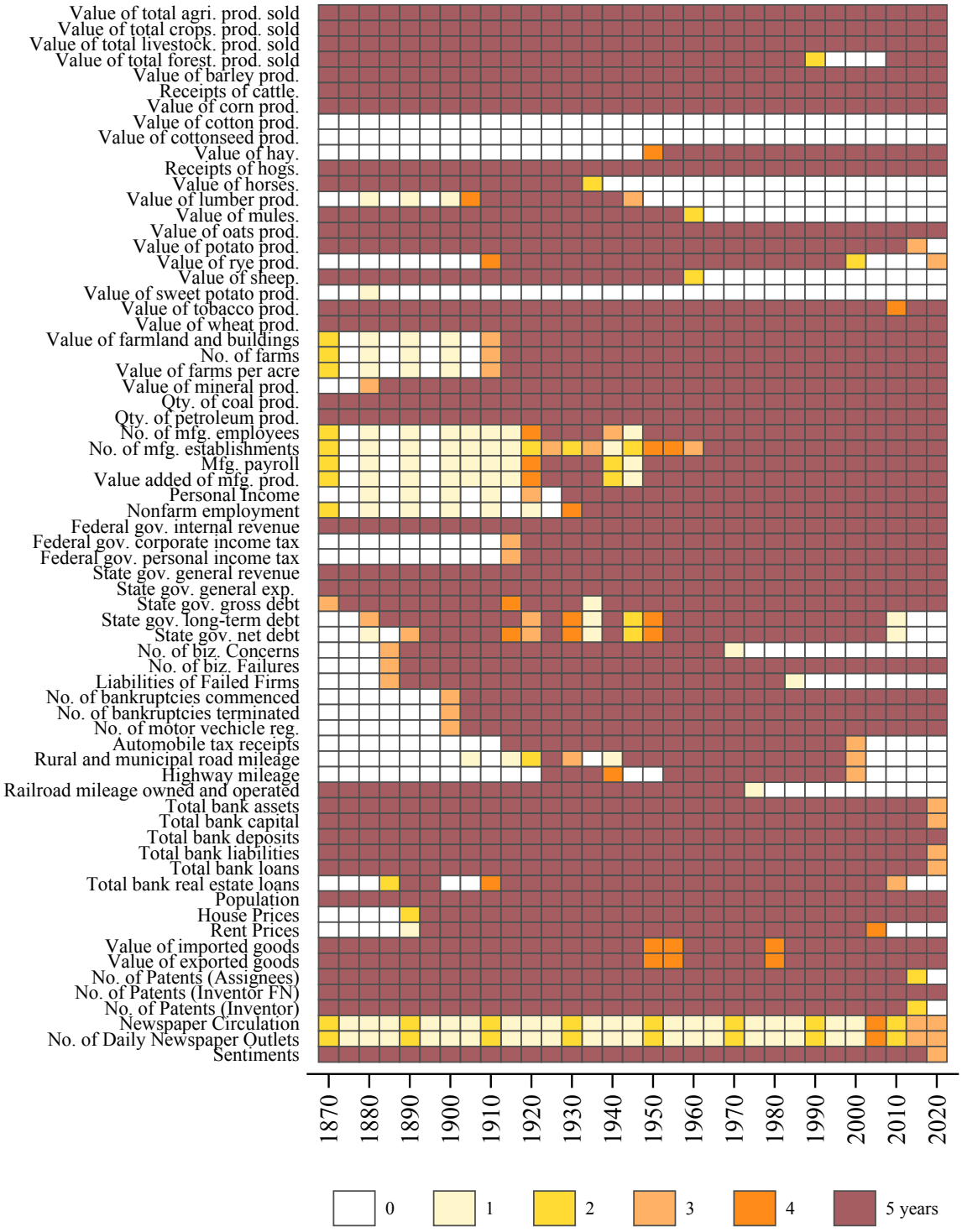
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 38: Availability of Variables – Oregon



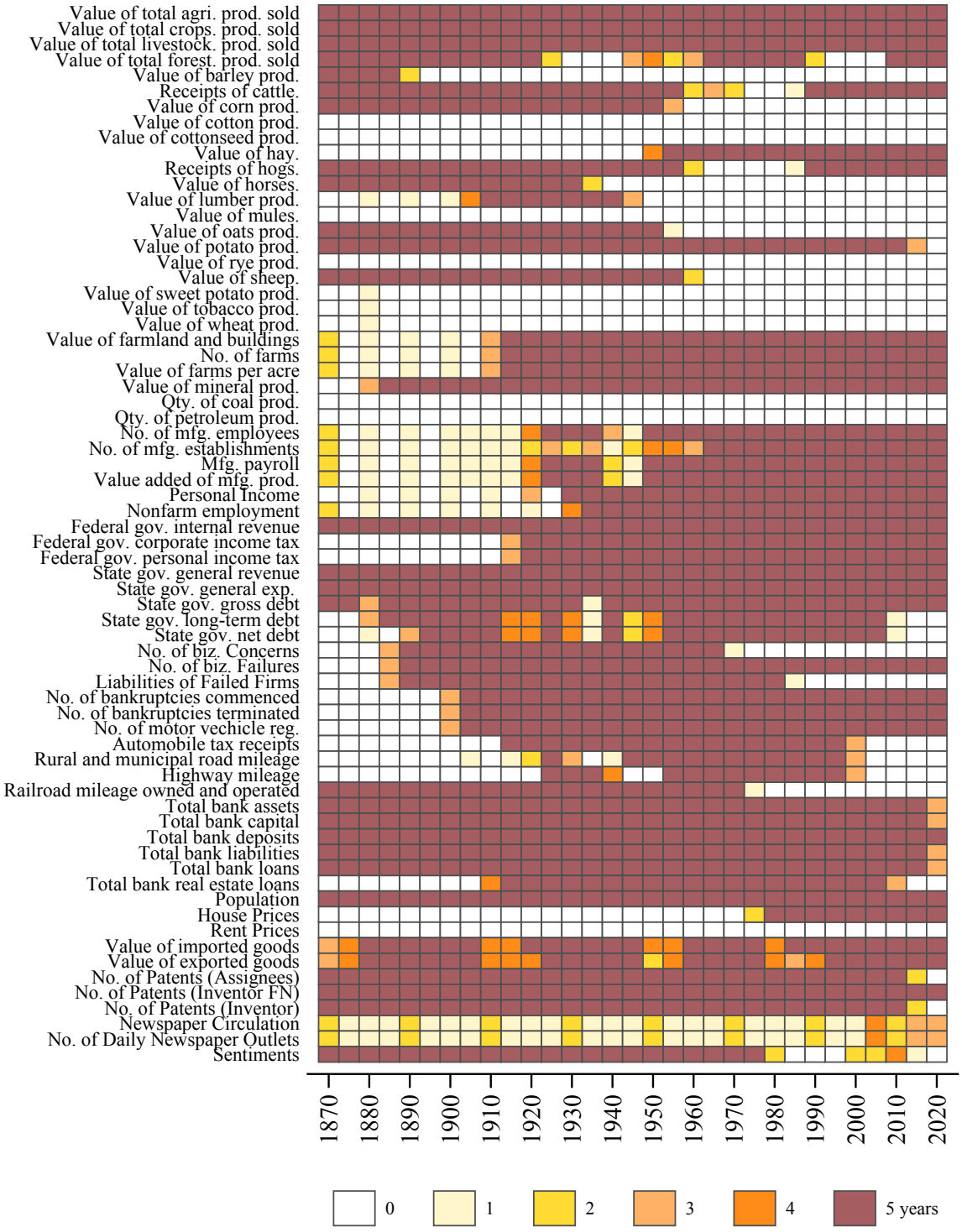
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 39: Availability of Variables – Pennsylvania



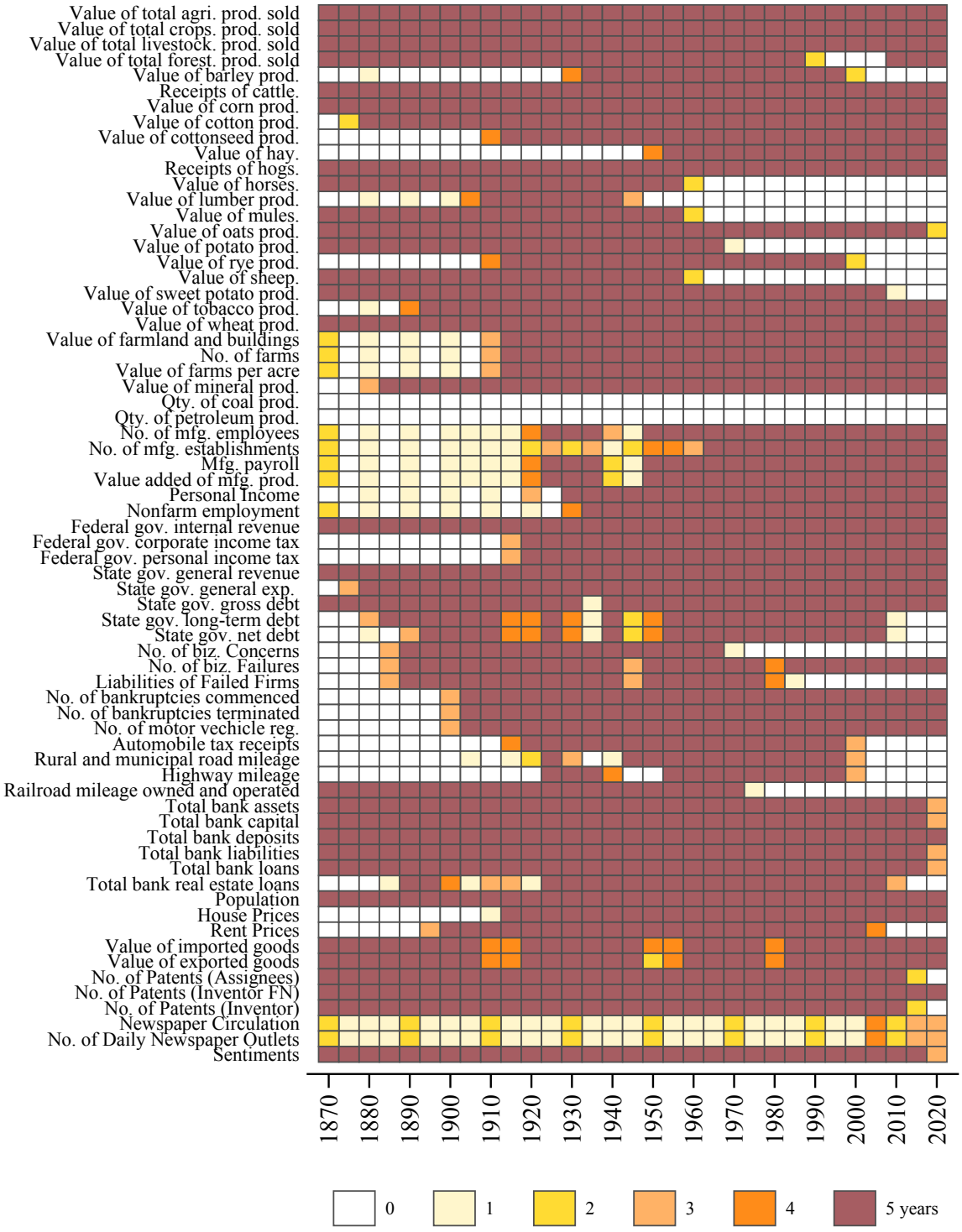
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 40: Availability of Variables – Rhode Island



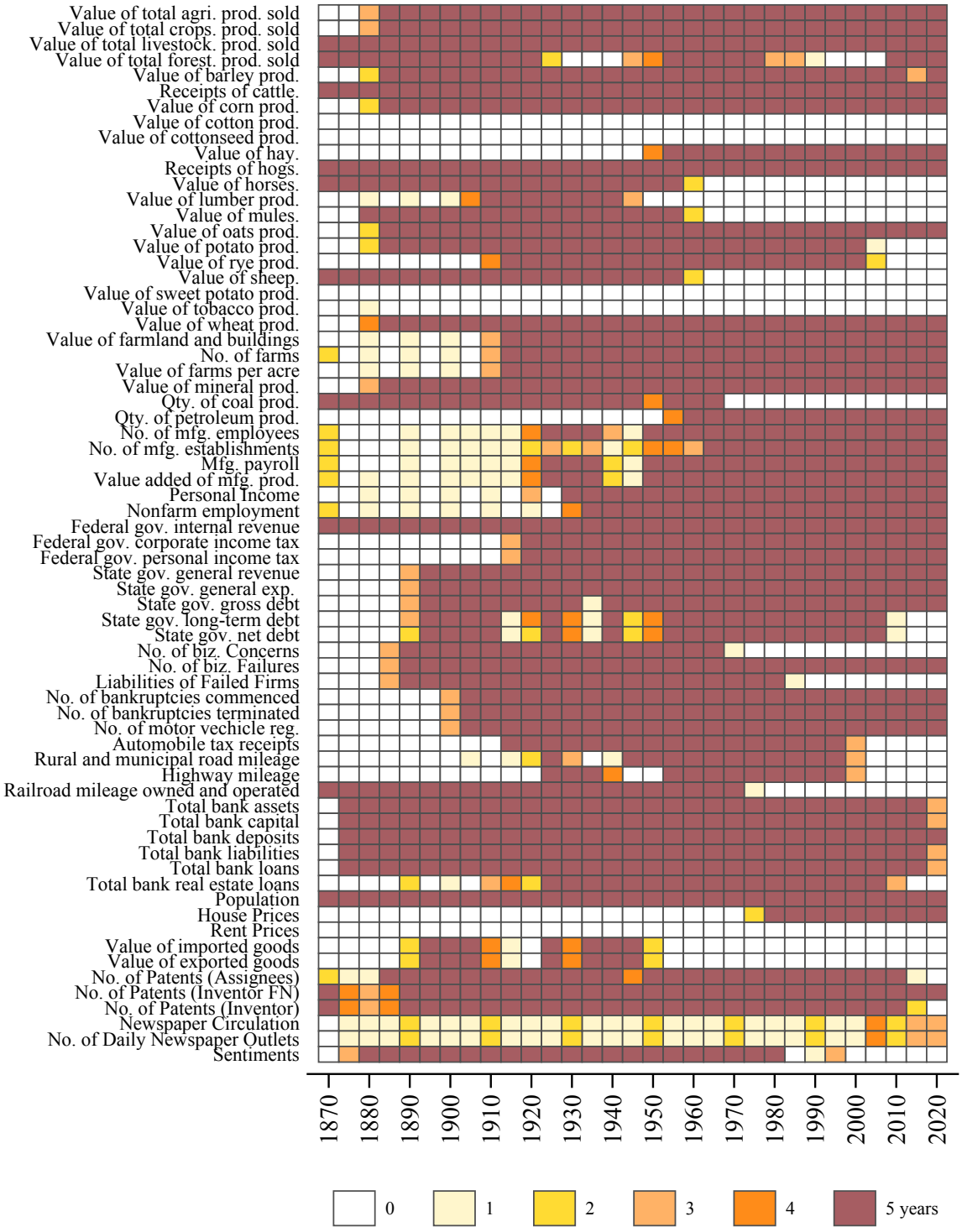
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 41: Availability of Variables – South Carolina



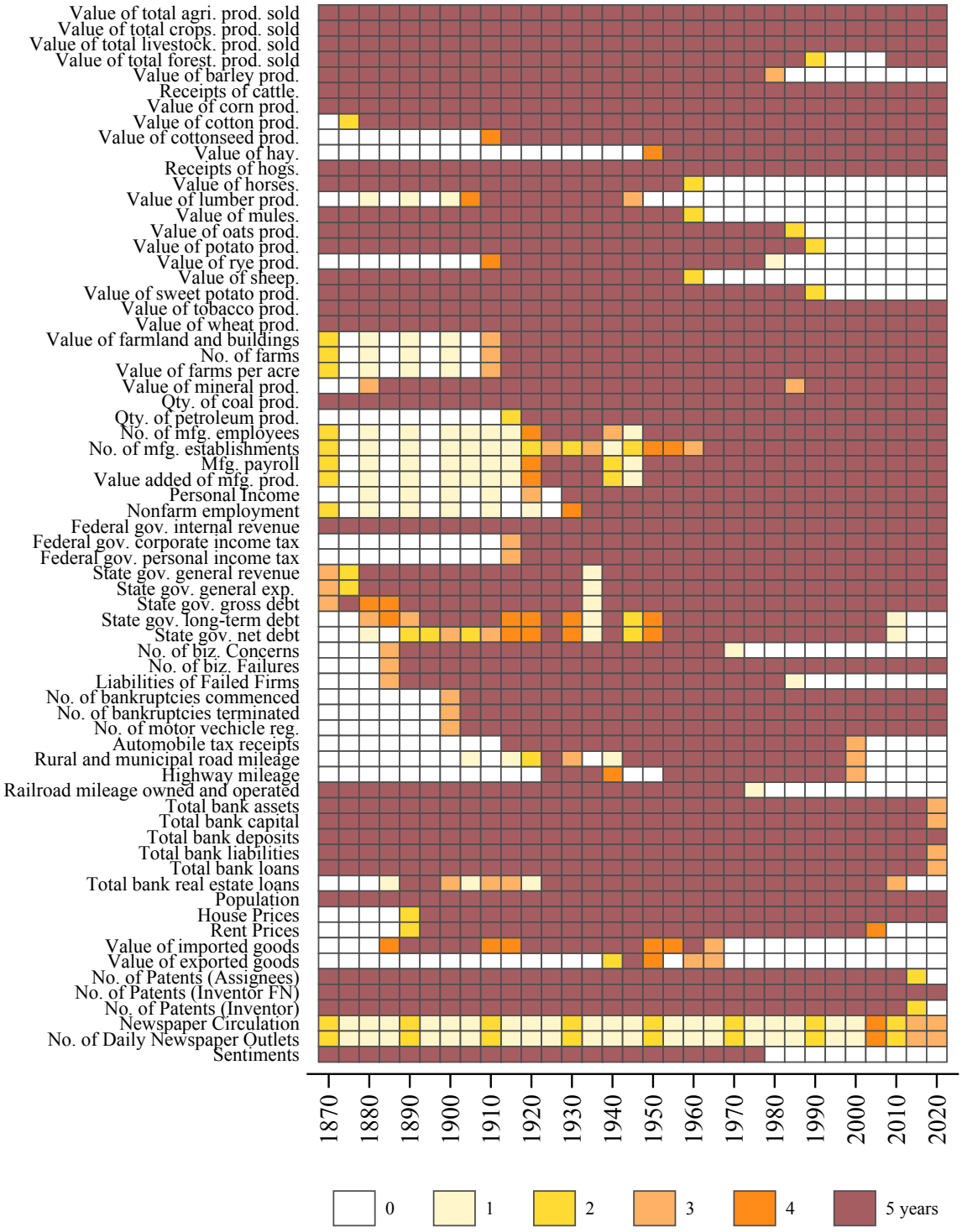
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 42: Availability of Variables – South Dakota



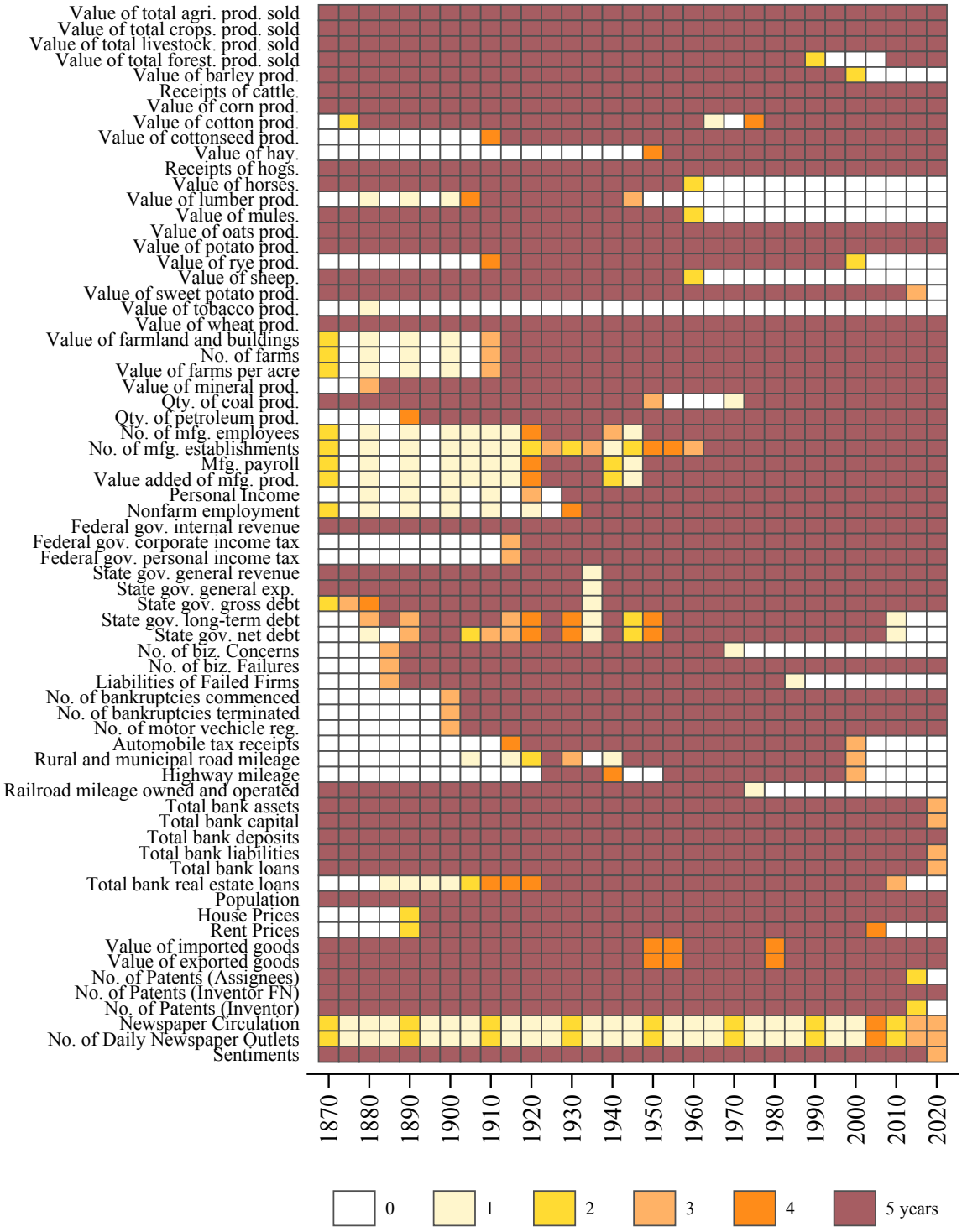
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 43: Availability of Variables – Tennessee



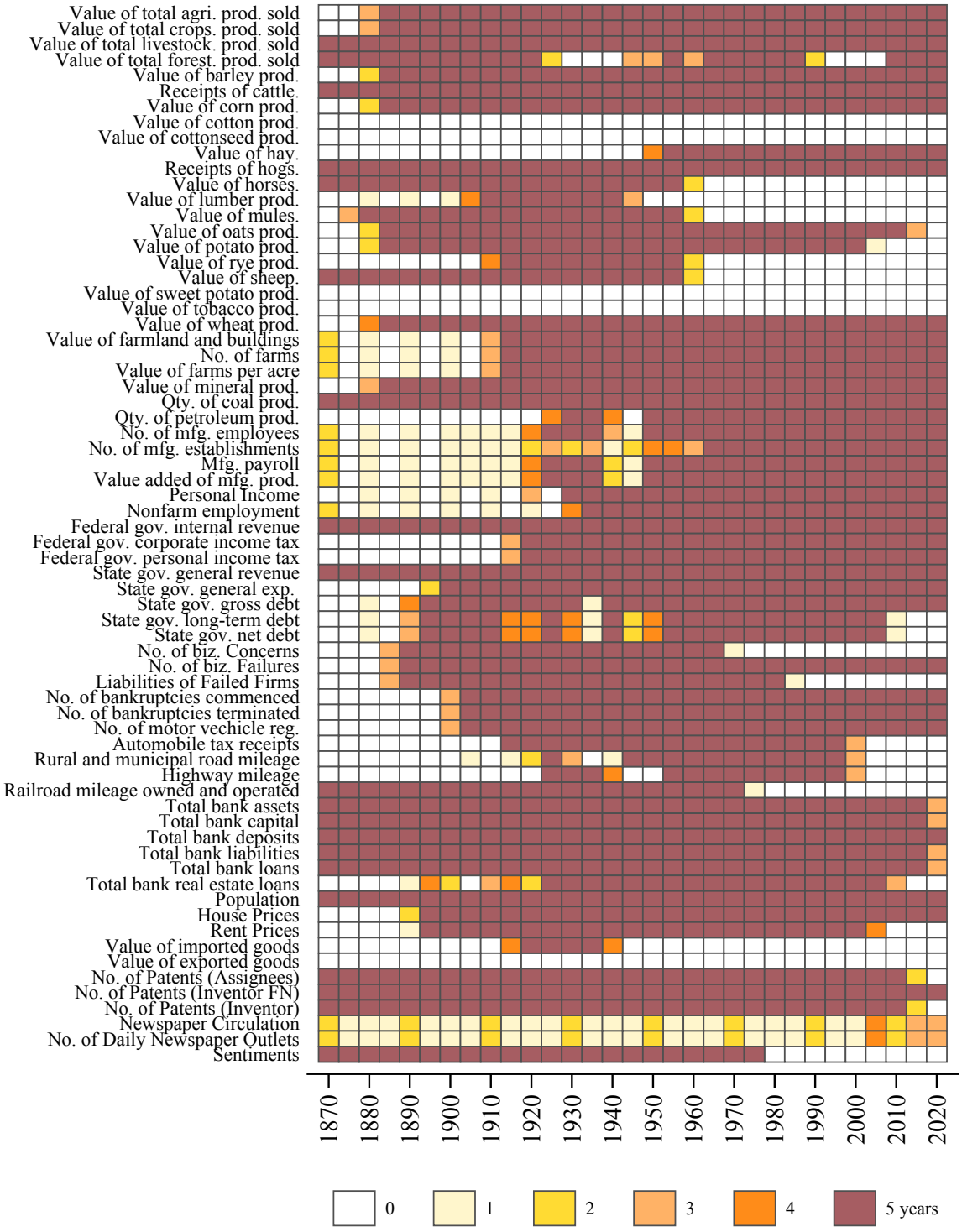
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 44: Availability of Variables – Texas



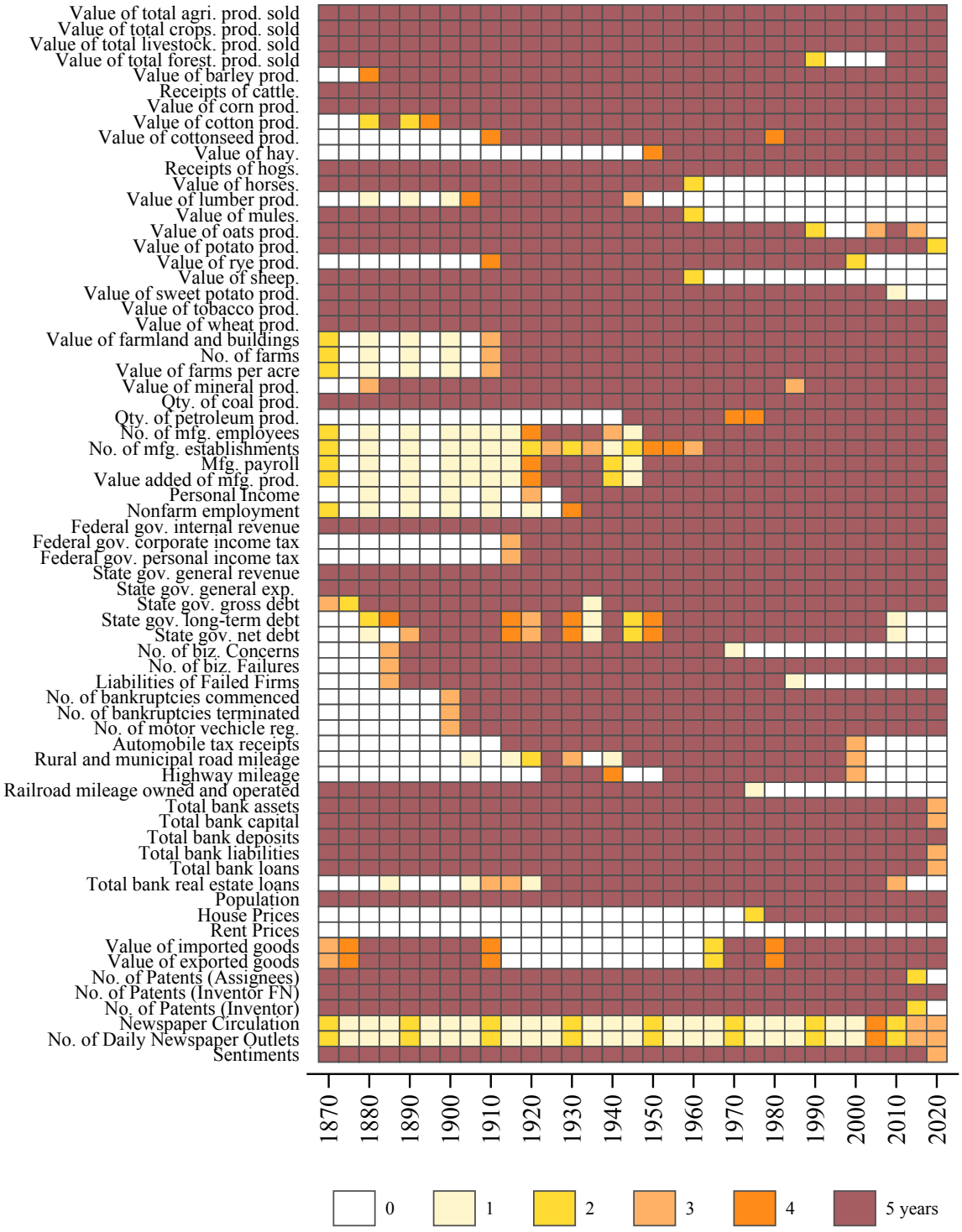
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 45: Availability of Variables – Utah



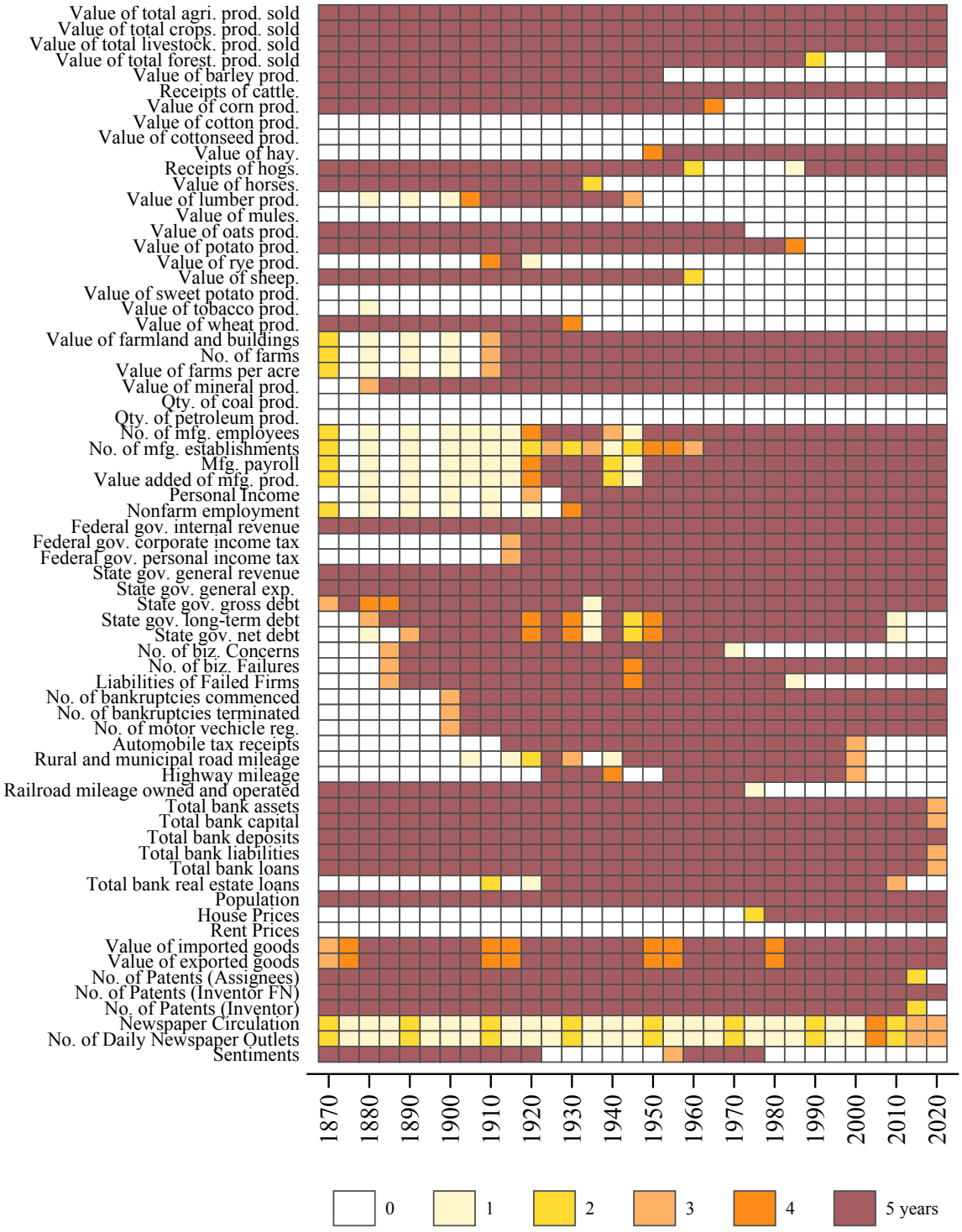
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 46: Availability of Variables – Virginia



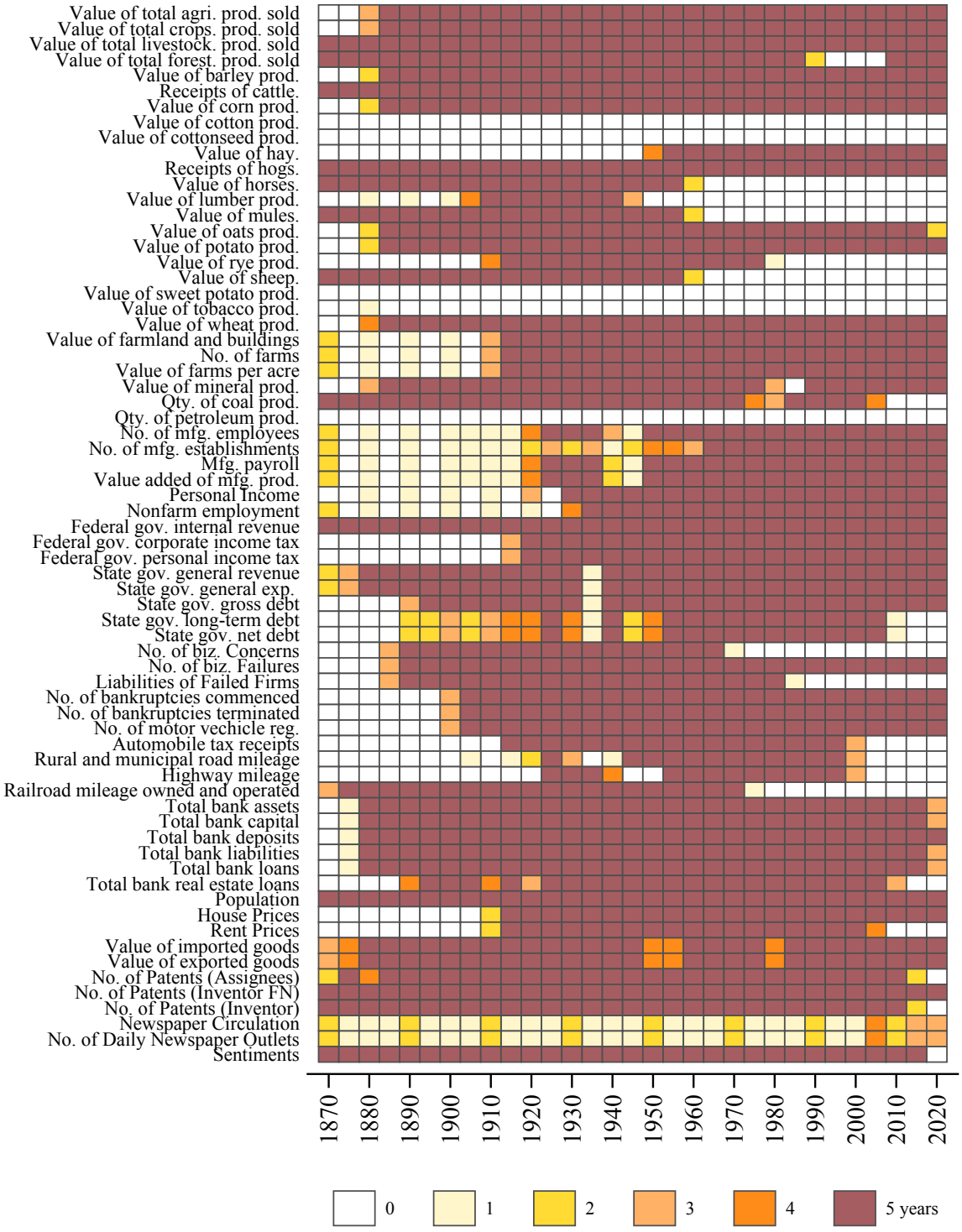
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 47: Availability of Variables – Vermont



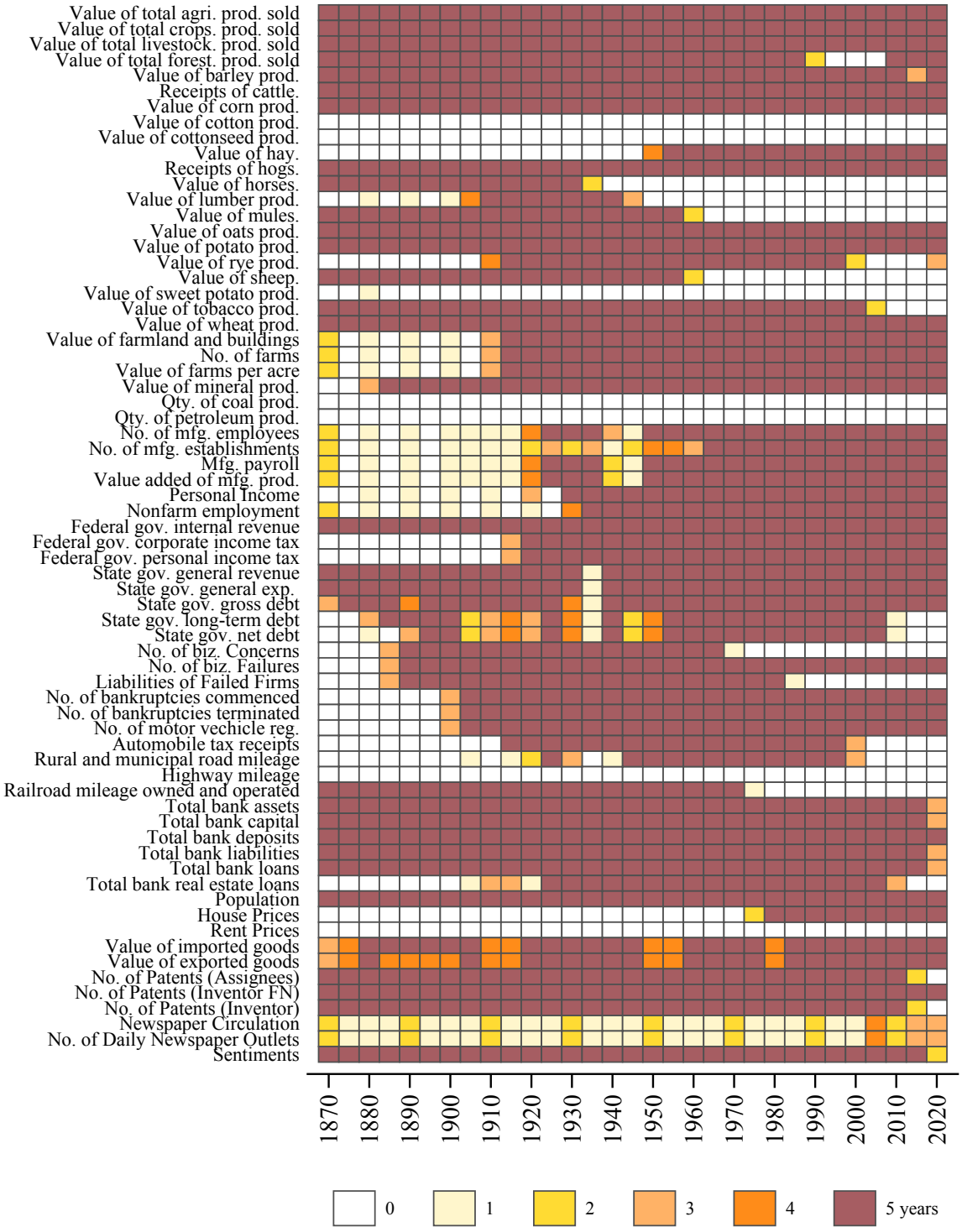
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 48: Availability of Variables – Washington



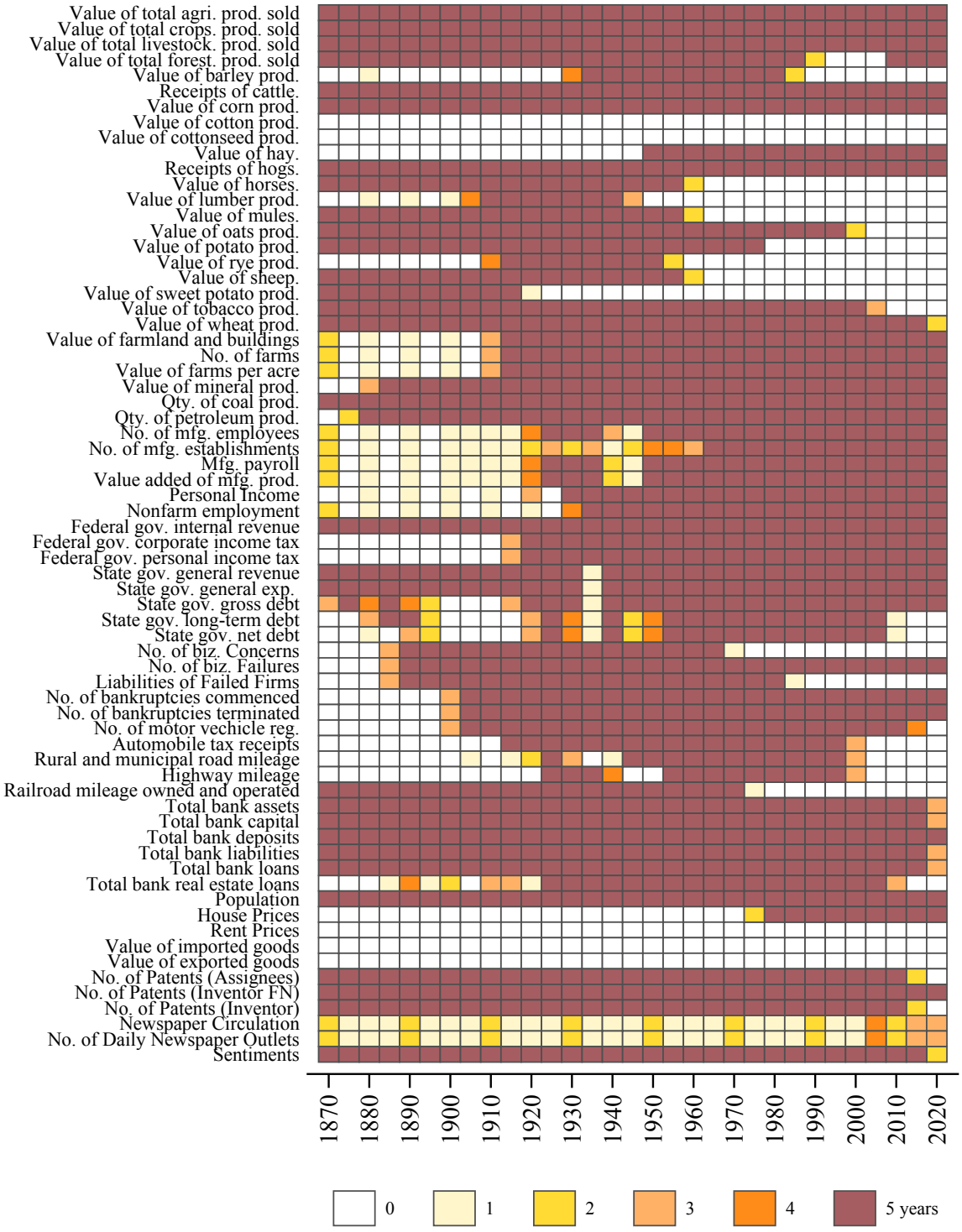
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 49: Availability of Variables – Wisconsin



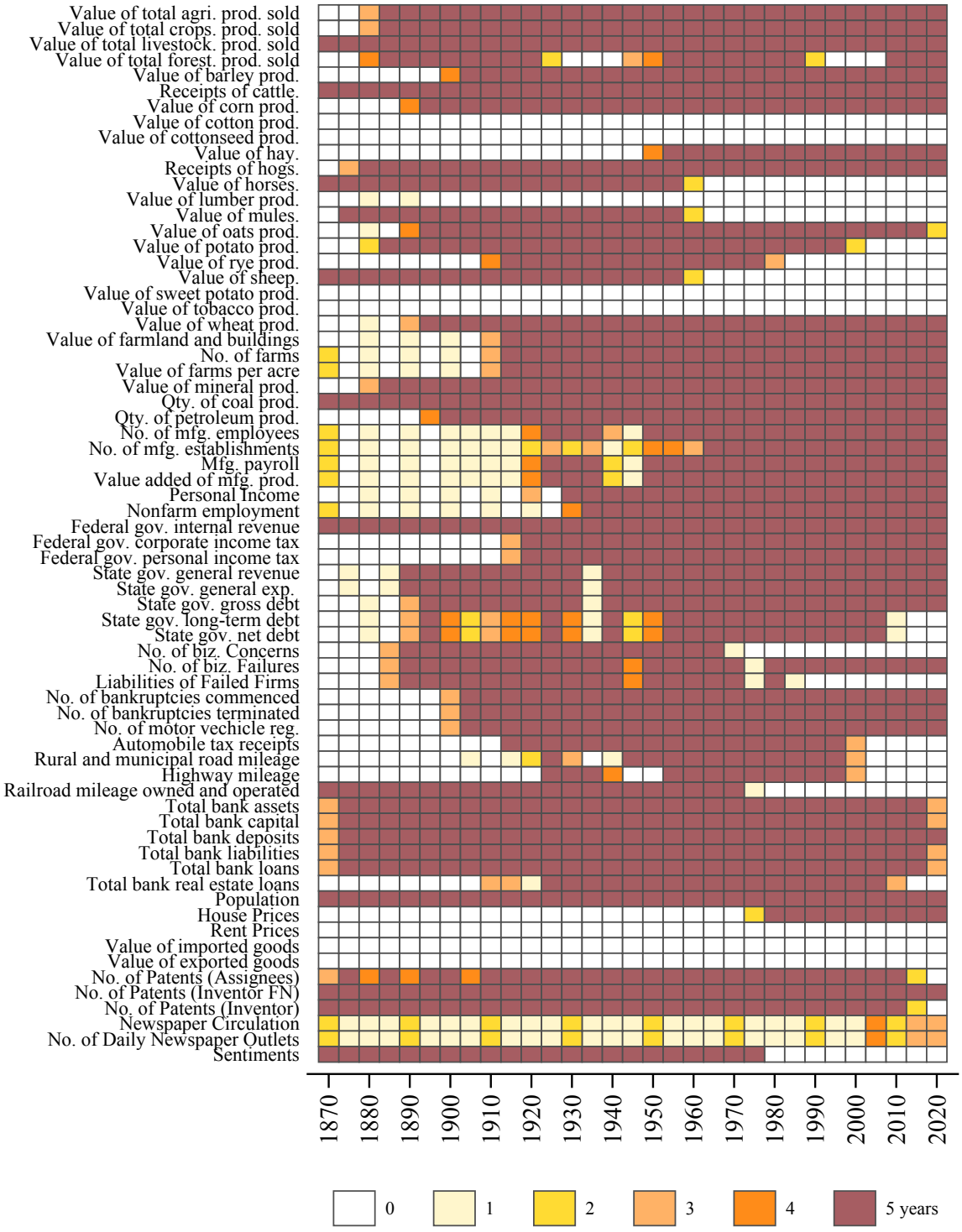
Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 50: Availability of Variables – West Virginia



Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.

Figure 51: Availability of Variables – Wyoming



Notes: This figure visualizes the availability of each variable within our dataset for the state in question, from 1870 to 2020. Each box represents a 5-year interval, with colors indicating the number of years for which data are available within that interval. The colors correspond to the following availability: 0, 1, 2, 3, 4, and 5 for data available in 1, 2, 3, 4, and 5 years within the interval, respectively.