The Emergence of a National Labor Market in US Manufacturing, 1900-1940

Matthew Hurt

January 14, 2022

1 Introduction

In 1912, the last of the lower 48 states were admitted to the Union. New technology made it easier for people to move to areas previously avoided by most migrants and sustain new communities there. It also made it easier than ever to move goods across the country. However, some elements of economic integration seemed to lag behind. Scholars like Rosenbloom (2002) have long argued that regional labor markets, sufficiently separated from a national labor market, persisted through this period and even up to and through the second World War. Studying how, why, and which labor markets remained isolated during this period helps us understand sources of push and pull forces responsible for increasing labor market integration in the United States.

This paper combines newly collected data from the US Census of Manufactures for 1900, 1910, 1920, 1930, and 1940 with a new measure of labor market integration to better measure the frictions present in American regional labor markets and their implications for overall regional labor market integration. In the past, scholars like Wright (1987) and Rosenbloom (1990) have focused on using data on real wages and an appeal to a law of one price to measure labor market integration. The law of one price is an economic theory that economic historians have borrowed from trade economists and while it is possible to implement it successfully in a historical setting as in Steinwender (2018), the lack of good historical data often make it a challenge to implement successfully.

Usually a law of one price is invoked to measure the size of bilateral frictions in a market for a homogeneous good with reliable price data. One way to utilize a law of one price to measure labor market integration involves comparing real wages for the same occupation across different regions of the country. The literature has found that real wages did converge over time, but that there were persistent differences in real wages between some regions such as in Wright (1987). In particular, Mitchener & McLean (1999) find that Southern wages tended to be lower than the national average while Western wages tended to be higher. Based on these differences in real wages, the literature (Wright 1986), Rosenbloom 2002), Collins & Wannamaker 2015) have inferred the presence of significant labor market frictions and additional evidence has led to a focus on the unique facets of the Southern economy and the conclusion that the South possessed a labor market

largely disconnected from the rest of the country.

However, there are two reasons to believe that the strategy of inferring labor market frictions from real wages is not appropriate in this setting. The first reason is that we do not have reliable individual wage data before it was reported in the US Census starting in 1940. Even in newer publications like Salisbury (2014) and Abramitzky et al (2017) authors that want to incorporate wage data into their analysis before 1940 either have to rely on scattered data or rely on wage proxies.

The second reason is that even if one can acquire enough wage data for a particular occupation or industry from a sufficiently large number of regions, the work that two individuals with the same occupation perform may be different enough to violate the law of one price assumption which requires the comparison of homogeneous labor market activity. Specifically, the concern is that even if we can control for region or industry specific characteristics there may be unobserved productivity differences that can bias estimation when using wages. Because of these limitations this paper employs a regression specification from Bernard et al (2013) that relies on using a ratio of wage bills for different types of labor within an industry-region pair as the outcome variable in a law of one price regression rather than relying on incomplete wage data.

However, wage bills may vary in competitive labor markets due to differences in the initial distribution of population. As in Bernard et al (2013) this paper uses a ratio of the wage bill for two types of workers: wage earners and salaried workers as the outcome variable to solve this problem. When using the ratio of two types of workers any differences in endowed population at the region industry level that would make a wage bill inappropriate to use for a law of one price regression are cancelled. For ease of exposition this ratio is referred to as a measure of skill premium since it measures the relative value each region industry pair places on employing highly skilled (salaried) or low skilled (wage) labor. Because these wage bills are the product of labor demand and wages unobserved productivity for each worker type can be controlled for.

As with real wages, once the industrial composition within a state is controlled for, the law of one price argument states that any differences we observe in average skill premium across regions

are the result of labor market frictions since different types of workers would migrate to secure higher earnings in the absence of labor market frictions. In other words, once we control for the fact that Michigan had a larger share of its industrial composition in high skill premium industries like automobiles, and Wisconsin had a larger share of its industrial composition in lower skill premium industries like dairy, differences in state level premium come from labor market frictions that prevent worker migration between skill levels, industries, and locations.

The wage bill data by region and industry come from the US Census of Manufactures. While this paper is not the first to use these wage bill data, to my knowledge, this is the first paper to use data for the decadal years 1900-1940 inclusive that includes all of the recorded industrial activity instead of just focusing on the largest industries. These data were hand recorded from scanned versions of the Census of Manufactures that are available for public access through the Census website. The data collected include all of the lower 48 states and the District of Columbia for 1900, 1910, 1920, 1930, and 1940 along with a panel of 44 cities described in the data section.

These new data and the new measure used in this paper contribute directly to the literature on postbellum labor market integration in the US. However, these contributions are also relevant for the literature that has measured regional integration such as Kim (1998) that measure regional integration through non-wage channels. Other papers like Ferrie (2005) that have studied labor market integration with respect to different groups of workers, such as immigrants, can also benefit from these new data as controls to better understand push and pull forces that affect workers decisions to relocate into different locations or occupations.

With a simple OLS regression of the log of the wage bills against a full set of industry and state fixed effects this paper finds two results about labor market integration in the United States. The first result is that labor market integration increased across the entire country between 1900 and 1940 as measured by the distribution of region fixed effects. This result is consistent with the existing narrative of labor market integration during this period. However, these increase was not smooth and it appears that the increase in labor market integration took place between 1920 and 1930. The second result is that there is weak, if any, evidence of Southern labor markets being

statistically different from any other regions. Neither the distribution, statistical significance, or magnitude of the region fixed effects support the idea that Southern labor markets included more labor market frictions than the rest of the country.

The conclusions of this paper do not seek to overturn our understanding that there were differences in regional labor markets. Meaningful differences existed in 1900 and continued to exist in 1940. However, neighboring labor markets having different wages or labor force composition by worker type are not necessarily evidence of poor labor market integration. If there are differences in regional preferences migration need not exist in the wake of wage differentials. This paper reinforces our understanding of important regional labor market differences, but also highlights the need to exercise caution when attributing the additional label of poor labor market integration to any one region of the American economy.

2 Background

The first half of the 20th century is often described as an era in which political, economic, and technological forces promoted economic integration in the United States. This period played host to several macroeconomic events such as the Panic of 1907 that led to the creation of the Federal Reserve (Friedman and Schwartz 1972), the first World War, the Great Depression and New Deal (Kennedy 1999) which all radically altered the economic landscape of not only the United States. There were also important changes in American labor markets caused by post war border closures in the 1920s studied in Abramitzky et al (2019) as well as the beginnings of the Great Migration as described by Collins (2021).

Rosenbloom (2002) observed that while goods market and political integration increased during this period, labor market integration lagged behind. In particular, they highlight differences between Northern and Southern labor markets exemplified by persistent gaps in real wages and a lack of expected migration as evidence of barriers to labor market integration. Using full count Census data for 1850-1940 from IPUMS (Ruggles et al 2019) researchers can verify that rates of migration

between the North and South were low before the Civil War, even lower after, and only after 1920 do we observe signs of a significant migration of Southerners to the North. Here the South is defined following Census guidelines and the North is defined as not the South.

To augment and provide context for the primary analysis this section documents three key facts about Southern and national labor markets in this period. The first fact is that Southern labor markets differed from other parts of the country. The second fact is that migration between the South and rest of the United States was relatively smaller than migration within both sub regions. Both of these facts are well established in the existing literature. This paper adds to the first fact by demonstrating that differences in wage patterns are present using the new more comprehensive Census of Manufactures data. This paper adds to the second fact by using full count US census data to track migrants between each of the 1900 - 1940 decadal censuses. The third fact highlights the importance of studying changes in salaried work using both full count census data and the new Census of Manufactures data.

Table 1: Division Level Wages Relative to the National Average

Census Division	1900	1910	1920	1930	1940
New England	95.86	97.99	91.85	93.23	93.65
Middle Atlantic	108.13	103.44	104.20	107.82	104.55
East North Central	101.59	105.76	107.89	112.27	117.45
West North Central	104.87	105.88	94.46	94.39	95.39
South Atlantic	71.53	71.29	82.69	68.77	69.86
East South Central	81.01	74.49	75.80	70.10	60.86
West South Central	90.05	92.45	88.77	80.73	79.79
Mountain	142.77	145.90	112.26	107.08	96.93
Pacific	116.38	140.53	115.42	108.09	119.13
Grand Average	101.35	104.19	97.04	93.61	93.07
Standard Deviation	19.61	24.05	12.90	15.91	18.79

Note: This table presents the average division level wage relative to the average national wage for each census division using data from the US Census of Manufactures. The grand average is an average of each element of the nine rows while the standard deviation is the standard deviation of the elements in the nine rows.

is a recreation of the first panel of table one from Mitchener and McLean (1999) which uses information from the US Census of Manufactures to show differences in the nominal wage of the

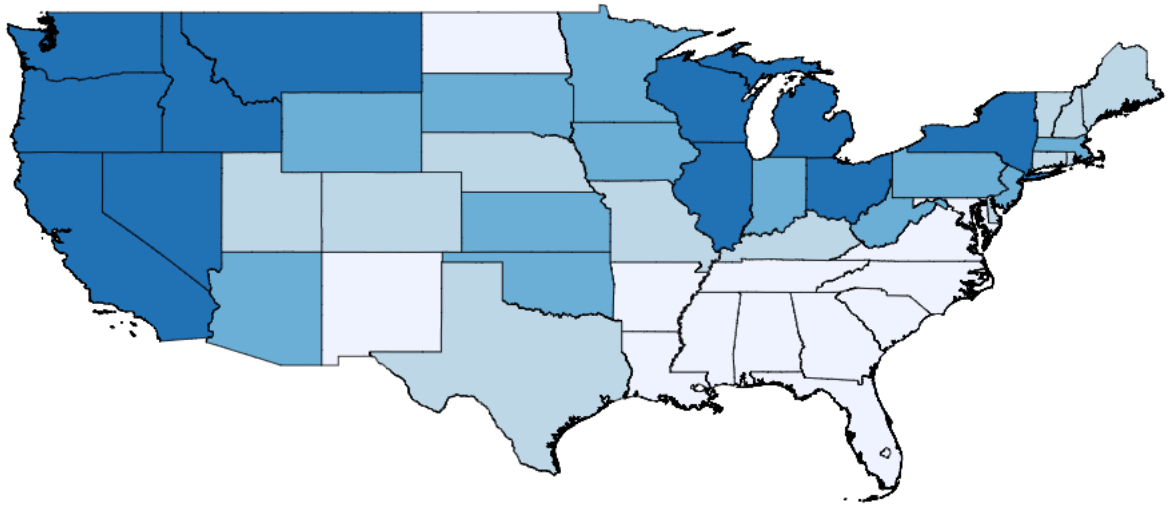


Figure 1: This figure shows the state level average wage computed using the US Census of Manufactures. The states are grouped by quartile with the darkest quartile including states with the highest average wages.

nine census divisions relative to the national average. Division wages were computed by summing the total wage bill for wage earners in a division and dividing it by the sum of of wage earners in that division. Here the data show that the three Southern census divisions in all five years had lower average wages than the other divisions. A rough potential measure of wage convergence is the standard deviation of each deviation. Across fifty years the standard deviation remained fairly consistent.[1]

Figure 1 is a map of the continental United States with state level wages sorted into quartiles with darker colors corresponding to the states with the highest wages. The state level wages were computed the same way the division wages were. While the map only presents the data for 1940 this is because there was little interesting variation in wages across states between 1900 and 1940. Between 1900-1940 Southern wage earners employed in industrial work consistently earned some of

[1]The New England division includes Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. The Middle Atlantic division includes New Jersey, New York, and Pennsylvania. The East North Central division includes Indiana, Illinois, Michigan, Ohio, and Wisconsin. The West North Central division includes Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota. The South Atlantic division includes Delaware, D.C., Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, and West Virginia. The East South Central division includes Alabama, Kentucky, Mississippi, and Tennessee. The West South Central division includes Arkansas, Louisiana, Oklahoma, and Texas. The Mountain division includes Arizona, Colorado, Idaho, New Mexico, Montana, Utah, Nevada, and Wyoming. The Pacific division includes Alaska, California, Hawaii, Oregon, and Washington, but in this paper excludes Alaska and Hawaii.

the lowest wages in the country. However, other parts of country, in particular the Pacific divisions, had wages that were persistently in the highest quartile.

For example, restricting the Census to only include men age 15-65 with known occupations who resided in the continental United States, shows that in 1900 over 97% of those born in the North were working in the North while 91% of those born in the South were working in the South, and 94% of all immigrants worked in the North. Between 1900-1940 the numbers for Northerners and immigrants did not change. Neither did the fact that the South consistently comprised 30% of the total labor force. What did change was the share of those born in the South working in the South. In 1920 the share fell to 89% and continued to fall to 85% by 1940. While seemingly modest these migrations from South to North represented the movement of millions of workers usually called the Great Migration.

There were also important geographic redistributions of the labor force between regions. Between 1900 and 1940 the Pacific Census division doubled its share of the national labor force going from 4% to 8% reflecting the beginnings of mass migration westward. The old frontier region of the West North Central saw its share of the labor force fall four percentage points while other parts of the North remained relatively steady. Within the South the East South Central lost roughly the same share of labor force that the West South Central gained. Based on these numbers one could conclude that any labor market redistributions were kept within the North and South. Given the persistent differences in wages, the corresponding lack of migration is rightly suspect.

However, incorporating recent linked Census data made available through the Census Linking Project from Abramitzky et al (2020) shows that there was substantial heterogeneity in the movement of workers between census divisions. For example, between 1900 and 1910 workers who left the West North Central were as likely to travel to the West South Central as they were all of the New England, Middle Atlantic, South Atlantic and East South Central census divisions. These kind of continued migrations between census divisions were common and not noticeably small for Southern divisions. These trends highlighted by more granular geographies should give researchers pause when declaring blanket statements about migration between the North and South.

Lastly, the period between 1900-1940 saw a continued decline in agricultural employment and increase in the importance of salaried industrial work. Figure 2 shows how the five major occupation groups defined using US census occupation descriptions evolved between 1850 and 1940. Starting in 1900 the decline in agriculture accelerates while the increase in white collar work accelerates rapidly. This white collar group includes, among other types of work, salaried workers in industrial employment. For a class of worker included in the largest employment group in '1940 and second largest between 1920 and 1930 the lack of analysis of salaried work has greatly limited economic historians' ability to analyze the industrial labor market.

Table 2: Division Level Salaries Relative to the National Average

Census Division	1900	1910	1920	1930	1940
New England	104.60	109.05	101.81	99.36	88.32
Middle Atlantic	105.38	102.54	103.45	104.28	103.40
East North Central	98.47	97.74	100.31	102.20	107.57
West North Central	93.90	92.25	87.95	87.44	88.71
South Atlantic	84.02	92.57	100.15	90.25	89.98
East South Central	95.61	92.11	94.28	88.76	79.81
West South Central	81.69	91.52	91.48	87.34	95.13
Mountain	105.95	110.62	95.09	85.52	90.88
Pacific	105.08	106.42	97.92	106.59	105.40
Grand Average	97.19	99.42	96.94	94.64	94.36
Standard Deviation	8.75	7.41	4.83	7.87	8.76

Note: This table presents the average division level salary relative to the average national salary for each census division using data from the US Census of Manufactures. The grand average is an average of each element of the nine rows while the standard deviation is the standard deviation of the elements in the nine rows.

Table 2 presents division level salaries relative to the national average salary. Once again the three Southern census divisions persistently have below national average salaries, but their deviation from the national average is consistently less extreme and other divisions more frequently have average salaries below the national average. Comparing the standard deviation of each average there is not evidence of salary convergence, but there is less variation in each year than there was for wages. This lower level of variation was driven by the three Southern divisions and two West divisions being closer to the national average.

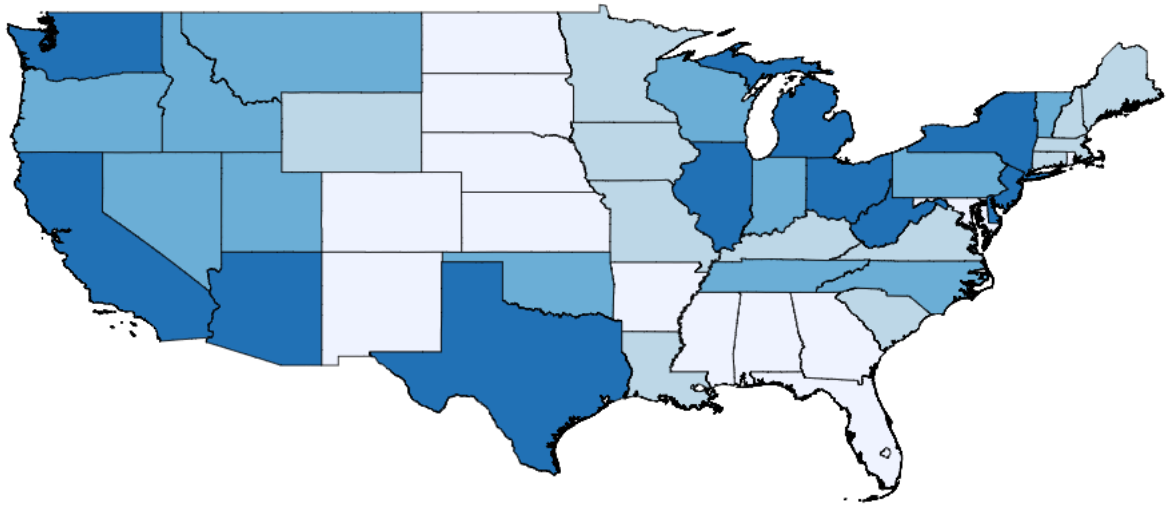


Figure 2: This figure shows the state level average salary computed using the US Census of Manufactures. The states are grouped by quartile with the darkest quartile including states with the highest average salaries.

Figure 2 demonstrates the greater degree of within region heterogeneity of average salaries along with a higher degree of heterogeneity across the entire US. Most, but not all of the states in the highest quartile for wages were also in the highest quartile for salaries. If quality price indexes existed at the state level for this time period real salaries may have been even similar across the country.

Together these three results add to and verify our existing knowledge of American regional labor markets. There were differences across space and these differences persisted through the period studied even as migration responses were slow to develop. However, new data show an additional layer of nuance to these trends as shown by close migration ties between adjacent regional labor markets that existed as early as 1900 and held throughout the period studied. Finally, new information about salaried work shows the need to study these group of workers who role in the economy only increased between 1900 and 1940 and who show signs of greater labor market integration than wage earners.

3 Empirical Methodology

Broadly speaking, a labor market is well integrated with other labor markets when workers can move between the labor markets at low cost. Since it is difficult to measure the costs associated with workers moving between labor markets, particularly with incomplete historical data, the literature has so far measured labor market integration by proxy. The most common approach has been to make an appeal to a law of one price to estimate labor market frictions by comparing real wages and the movement of workers between labor markets.

The idea of a law of one price comes from the trade literature and argues that, given a specified functional form for trade, if two regions produce an identical good, prices should converge as trade barriers fall. In a case with no trade barriers and completely identical goods product markets would be perfectly competitive. In practice, the law of one price does not seem to hold empirically as seen in work like McCallum (1995) and Anderson & Van Wincoop (2003), but it can still be a useful tool for measuring trade frictions. If the same kind of logic is appropriate to use when studying labor markets, then researchers can infer labor market frictions by measuring deviations from a law of one price in real wages.

As shown in the previous section, wages were persistently lower in the the South and migration between the South and rest of the country was rare relative to migration within the South and within the rest of the country. In the context of a law of one price, this can only hold in the presence of labor market frictions. However, in the context of labor markets, a law of one price demands that all workers be identical. Scholars have tried to work around this assumption by comparing real instead of nominal wages, but the data used for their price indexes and wages were incomplete. Furthermore, even if the data were good, using real wages at best controls for differences between regions, but fails to address any observable or unobservable differences in productivity between workers in different industries and between worker types within an industry.

A better strategy compares the ratio of wages of two arbitrary types within each possible region-industry pair. Equation (1) presents an example of how a regression that uses such an outcome variable can be used to measure variation in labor market frictions across different regions and

industry groups. Previously, the requirement of the law of one price was that as labor market frictions decline, wages for all workers should converge. A weaker requirement would be for wages to converge within worker types as labor market frictions decline. If the numerator and denominator of a fraction are converging over time their fraction should as well so this wage ratio should also converge as frictions decline.

$$\frac{wage^{a}_{rjt}}{wage^{b}_{rjt}} = \alpha_{rt} + \mu_{jt} + \epsilon_{rjt} \tag{1}$$

In this equation, workers belong to either type a or b. The different regions (states) are indexed with r, while different industries are indexed with j. The advantage of using a ratio of wages is that any observable terms that are common to workers of both types in each region-industry pair, such as price levels, will be cancelled out. Furthermore, any unobservable components of wages common to workers of both types within a region-industry pair will cancel out as well.

In a regression like this the α_{rt} are estimated using variation in wage ratios across all industries within region r. As labor market frictions decrease, the α_{rt} should converge. Likewise, the μ_{jt} are estimated using variation in wage ratios across all regions within a particular industry j. Changes in labor market frictions are not expected to have any effect on the μ_{jt}. Instead the μ_{jt} are likely more dependent on the ratio of worker types needed for a particular industry. For example, in some industries it might be important to get the best and most expensive workers of type a, but workers of type b may perform a task that can be completed by any worker.

While this is a significantly better way to measure labor market frictions than what has been done previously, there is one way to improve on this technique without requiring a structural model or additional data. Other research such as in Bertrand et al (2013) show that using wage bills instead of wages allows researchers to not only control for observable and unobservable productivity differences common to *all workers* in each region-industry pair, but also control for unobservable productivity differences common to workers of *each type* within each region-industry pair.

In the context of this paper, a wage bill for each worker type is the total expenditure paid by firms in each region-industry pair to each type of worker. In other words, a wage bill is the product

of labor demand and wages. Like with the analysis of efficiency units in Trefler (1993) first define the productivity adjusted labor demand for workers of type a in some region-industry pair x^a_{rj} as the product of an unobserved productivity component θ^a_{rj} and the observed labor demand $\tilde{x}^a_{rj}$.

$$x^a_{rj} = \theta^a_{rj}\tilde{x}^a_{rj} \tag{2}$$

Next define the wage rate firms that employ $\tilde{x}^a_{rj}$ as $\tilde{w}^a_{rj}$ and the wage rate firms pay for x^a_{rj} as w^a_{rj}. If firms hire one unit of $\tilde{x}^a_{rj}$ at $\tilde{w}^a_{rj}$ they receive effective productivity of $\tilde{x}^a_{rj}$, which means that firms that only want one unit of productivity need only hire $\frac{1}{\theta^a_{rj}}$ units of x^a_{rj} priced at $w^a_{rj} = \frac{\tilde{w}^a_{rj}}{\theta^a_{rj}}$

$$w^a_{rj} = \frac{\tilde{w}^a_{rj}}{\theta^a_{rj}} \tag{3}$$

Given this construction, a wage bill $x^a_{rj}w^a_{rj}$ completely cancels out the unobserved productivity parameter common to each worker type within each region industry pair. It also means that the observed wage bill is exactly equivalent to the productivity adjusted wage bill. The same argument that a ratio of wages between worker types should converge as labor market frictions decline applies to the argument that a ratio of wage bills between worker types should converge as labor market frictions decline provided that the convergence is not driven by a convergence in the size of the respective labor forces. Appendix Table C shows a kernel density plot of the coefficients from equation (4) estimated with the log of salaried workers to wage earners. The plot shows that there was no convergence over time in relative employment patterns.

$$ln(\frac{wagebill^S_{rjt}}{wagebill^{NS}_{rjt}}) = \alpha_{rt} + \mu_{jt} + \epsilon_{rjt} \tag{4}$$

Equation (4) is the primary estimation strategy used in this paper. Now the group types consist of salaried workers and non-salaried workers(wage earners). While the groups can theoretically be arbitary, the background section demonstrated the importance of including salaried workers in this kind of labor market analysis. It also provides a convenient proxy to contemporary analysis of differences between high and low skilled workers. The outcome variable is taken in logs instead of

levels to remove any unobserved region-industry characteristics that only have proportionate effects on wages and employment.

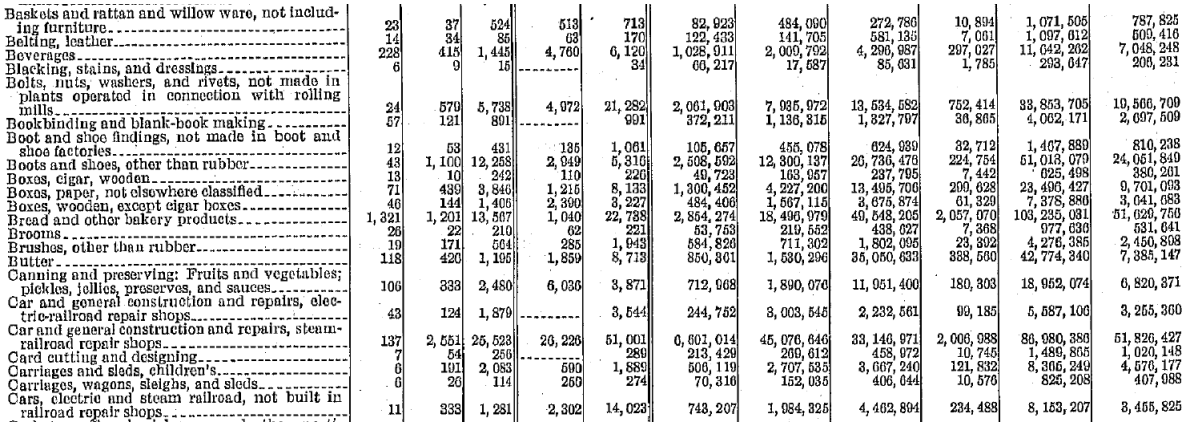
The interpretations of the fixed effects are similar to the interpretations of (1) only now this wage bill ratio should be thought of as a measure of the skill premium each region-industry pair pays out. In other words, the variation used to estimate α_{rt} come from the fact that in some regions the total expenditure for salaried workers across all industries is relatively greater than the total expenditure for wage earners. This may be because of the previous example detailed for the wage ratio, but may also be because certain regions are trying to attract a relatively larger amount of salaried workers even if wages for salaried workers are not relatively high.

Lastly, if one objects to the notion that the use of wage bills completely controls for the unobserved productivity differences of each type of worker in each region-industry pair, the use of a wage bill ratio instead of a wage ratio should at least do no worse provided that the expected convergence in light of declining labor market frictions is not happening exclusively through convergence in labor demand. As constructed, any differences between θ_{rj}^{S} and θ_{rj}^{NS} are eliminated. A regression that fails to account for this difference will be less well estimated.

While equation (4) is the primary regression equation used in this paper, one can use equation (1) instead and find results that are broadly consistent, but appear to be attenuated and have larger standard errors. This outcome speaks to the need for some strategy to control for important differences in productivity at the worker type level, which wage bills theoretically can and empirically seem to do.

4 Data

To estimate equation (4) data on the total compensation paid by firms to workers earning wages or salaries that vary by region, industry, and year are needed. The data were hand recorded from the US Census of Manufactures for 1899, 1909, 1919, 1929, and 1939. For ease of exposition the years 1900, 1910, 1920, 1930, and 1940 are used respectively. Each year three sets of data covering



Baskets and rattan and willow ware, not including furniture	23	37	524	513	713	82,923	484,090	272,786	10,894	1,071,505	787,825
Belting, leather	14	34	85	63	170	122,433	141,705	581,136	7,061	1,097,612	509,416
Beverages	228	415	1,445	4,760	6,120	1,025,911	2,000,792	4,296,987	297,027	11,042,262	7,048,248
Blacking, stains, and dressings	6	9	15		34	66,217	17,587	85,631	1,785	293,047	203,231
Bolts, nuts, washers, and rivets, not made in plants operated in connection with rolling mills	24	579	5,738	4,072	21,282	2,061,903	7,935,972	13,534,582	752,414	33,853,705	19,566,709
Bookbinding and blank-book making	57	121	801		991	372,211	1,136,315	1,327,797	36,865	4,062,171	2,697,509
Boot and shoe findings, not made in boot and shoe factories	12	53	431	135	1,061	105,657	455,078	624,939	32,712	1,467,889	810,238
Boots and shoes, other than rubber	43	1,100	12,258	2,949	5,316	2,508,592	12,300,137	26,736,476	224,754	51,013,079	24,051,849
Boxes, cigar, wooden	13	10	242	110	226	49,723	163,957	237,795	7,442	625,498	380,261
Boxes, paper, not elsewhere classified	71	439	3,846	1,215	8,133	1,300,452	4,227,200	13,495,706	299,628	23,496,427	9,701,093
Boxes, wooden, except cigar boxes	46	144	1,405	2,390	3,227	484,406	1,667,115	3,675,874	61,329	7,378,886	3,041,683
Bread and other bakery products	1,321	1,201	13,567	1,040	22,788	2,854,274	18,496,979	49,548,205	2,057,070	103,235,031	51,629,756
Brooms	26	22	210	62	221	53,753	219,552	433,027	7,368	977,630	531,641
Brushes, other than rubber	19	171	564	285	1,943	584,826	711,302	1,802,095	23,302	4,276,385	2,450,808
Butter	118	420	1,195	1,859	8,713	850,361	1,530,296	35,050,633	388,560	42,774,340	7,385,147
Canning and preserving: Fruits and vegetables; pickles, jellies, preserves, and sauces	106	333	2,480	6,036	3,871	712,968	1,890,076	11,951,400	180,303	18,952,074	6,820,371
Car and general construction and repairs, electric-railroad repair shops	43	124	1,879		3,544	244,752	3,003,545	2,232,561	99,185	5,587,106	3,255,360
Car and general construction and repairs, steam-railroad repair shops	137	2,551	25,523	26,226	51,001	6,601,014	45,076,646	33,146,971	2,006,988	86,980,388	51,826,427
Card cutting and designing	7	54	256		289	213,429	269,612	458,972	10,745	1,489,865	1,020,148
Carriages and sleds, children's	6	191	2,083	590	1,889	506,119	2,707,535	3,667,240	121,832	8,305,249	4,576,177
Carriages, wagons, sleighs, and sleds	6	26	114	260	274	70,316	152,035	406,044	10,576	825,208	407,988
Cars, electric and steam railroad, not built in railroad repair shops	11	333	1,281	2,302	14,023	743,207	1,984,325	4,462,894	234,488	8,153,207	3,455,825

Figure 3: This is an excerpt from the 1929 Census of Manufactures that records industrial activity for a particular state. Each column represents a variable like the number of establishments, level of employment, wage bill, or value generated by the industry.

different regions are constructed. The first set includes the 48 continental states and the District of Columbia. The second set includes a panel of the 44 cities included in each of the five years of the Census of Manufactures. The final set includes the same 44 cities, and to avoid double counting, the 48 continental states minus any cities present in those states. For example, the third data set includes the state of Idaho, the city of Portland, Oregon and the state of Oregon minus the city of Portland, Oregon.

Figure 3 is an excerpt from the 1930 Census of Manufactures and shows the format of the acquired data. Different years included different information, but all five years included a description of the industrial activity, the number of establishments, the number of wage earners, the number of salaried workers, the wage bill of wage earners, the wage bill of salaried workers, and the value of each entry of industrial activity. Figure 1 also shows that sometimes data are omitted from the table. This is not because the true value is zero, but because the recording authorities either wanted to protect the anonymity of workers or deemed the amount of industrial activity insignificant.

Industrial activity were consolidated into a panel of 99 industry groups that were consistent across the five years. In all regression analysis the last industry group for other activity is excluded. These industry groups were based on SIC-3 groupings whenever possible, but there were three forces that resulted in a less granular grouping. The first is related to the development and obsolesce of certain kinds of industrial activity. For example, in 1900 there was no aircraft industry in the United

States, but after 1900 there were no entries for kaolin earths. Furthermore, OSHA last updated the SIC classification system in 1987 so many kinds of older industrial activity were left unclassified. In these cases industrial activity are grouped as similarly as possible while aiming for as many granular industry groups as possible.

The second force that resulted in less granular grouping is the fact that different years have different descriptions for the same industrial activity. For example, in 1920 the table reported separate production for cigars, cigarettes, chewing tobacco, and other tobacco production that matched existing three and four digit SIC codes. However, the 1940 census only reported activity in the tobacco industry that could be matched to a 2 digit SIC code. In this case, granularity was sacrificed for consistency through time by creating a separate group for all tobacco products in all years. Regardless of these discrepancies all industrial activity was included in one the final 99 industry groups. Even more confounding is the fact that cities and states with the same year used different descriptions for the same kind of industrial activity.

The final force has to do with the variation in granularity recorded by each state. The state level data included 5265 rows of activity for 1900, 2725 rows for 1910, 4670 rows for 1920, 4194 rows for 1930, and 6532 rows for 1940. Cities had fewer entries, but the patter across years was consistent. Creating industry groups that would have at least one element per state required consolidation in a way that resulted in some groups being very granular while others are quite broad. Appendix Table B lists the 99 industry groups. The list includes highly specific industry groups for activity such as $Coffins$ or $Brooms$, $Brushes$ while also including the more generic $Miscellaneous\ Metals$ or $Trains$, $Other\ Transportation$. In addition to the three sets described above robustness checks for data at the establishment level as well as state level analysis at the two digit SIC level are included to demonstrate the robustness of the regression strategy.

Table 3 shows how the number of industry groups varied across time at the state level. The first trend to highlight is that in no year did one state have every single industry group, but at least one state had one of each of the 99 industry groups. The second trend is that over time it was states with the least industrial activity whose variety in industrial activity declined over time.

This may be because the SIC 3 classification is too modern for the time period covered. Still, time trends in data classification are not a concern because all analysis is done separately for each year. The last trend to highlight is the variation of industrial composition throughout states. The number of industry groups is highly correlated with levels of employment and total labor earnings. Additionally, in all five years it is the Great Lakes states as well as states like Pennsylvania, Ohio, and New York that have the highest levels of industrial labor earnings and industrial variety.

Table 3: Industry Group Characteristics by State and Year

Year	1900	1910	1920	1930	1940
Industry Groups: Total Number	99	99	99	99	99
Industry Groups: Minimum per State	23	12	13	8	7
Industry Groups: Maximum per State	97	93	97	98	96
Industry Groups: Average per State	60	44	54	51	56

Note: This table presents information on the total number of distinct industry groups per year, the amount of industry groups in the state with the fewest, and the amount of industry groups in the state with the most. This data only includes industry group-state pairs with non-zero values for salaries and wages for that pair.

To determine which industries tended toward higher skill premia the wage bills for both types of workers across all states were summed for each year and then the log ratio was computed for each industry. The industries with the highest skill premia over time tended to ones like the medicine, publishing, and cleaning supplies industries. Industries that tended to have lower than average skill premia included clothing manufacturing in 1930 and 1940 as well as industries related to train maintenance and shipbuilding across all five years. A larger skill premium means that these industries put forth a larger ratio of labor payments to salaried workers instead of wage earners.

The 99 industry groups were designed around the state level data. Even so they do not form a balanced panel. Instead the industry groups were constructed to ensure that at least one state in every year included that industry group. Analysis at the city level did not include a reconfiguration of the industry groups so it is possible that in some years some of the industry groups are missing. Constructing the city level data followed the same process as the construction of the state level data. The panel of 44 cities detailed in Appendix Table A: Cities were selected solely because they were the only cities that were included in all five years. While city boundaries are more prone

to change than state boundaries are the selection of a panel of cities ensured as much geographic consistency as possible.

The last set of data includes the 44 cities as well as the states minus the cities. This third set of data provides more observations for statistical tests of similarity for each year. The decision to subtract cities from their relevant states is to avoid double counting. By construction this data set will include all 99 industry groups in all years. There were several instances, particularly in the 1939 data, in which the recorded industrial activity for cities exceeded the activity recorded by the states. Every instance was investigated and all cases were either genuine errors in the Census of Manufactures or the result of cities and states classifying different industrial activity into different groups. In all cases these negative values were zeroed out after being reviewed.

The state level data are the primary data set because it has the most consistent and representative geography. Using cities instead of states can be more attractive since most industrial activity took place within cities, but there is a selection issue regarding which cities are included in the panel. Only cities that had enough industrial activity between 1900 and 1940 to be included in the Census of Manufactures each year were included in the panel. This means that many cities with substantial industrial activity either early or later were excluded. Even without selection concerns, the cities are geographically concentrated in a way that is not representative of the entire United States. The last set is not preferred only because of the heightened risk for errors in the data after subtracting cities from relevant states.

5 Results

This section covers the results from estimating equation (4) using the hand collected Census of Manufactures data on state level industrial activity for 1900, 1910, 1920, 1930, and 1940. Recall that the main outcome variable α_r are state fixed effects that measure the difference between the average state skill premium and the national average skill premium were skill premium is defined as the log of the ratio of salaried wage bills to wage earner wage bills. In a well integrated labor market

we expect these state fixed effects to be statistically indistinct from zero. In other words, once we control for the industrial composition of each state, the average skill premium in each state should be statistically indistinguishable from the national average skill premium in a well integrated labor market.

Table 4: Main Regression Results Excerpt

Variable	Ratio				
Year	1900	1910	1920	1930	1940
Delaware	-0.05	-0.086	-0.074	-0.05	-0.089
	(0.142)	(0.115)	(0.092)	(0.112)	(0.077)
DC	-0.165	-0.374**	0.131	0.18**	0.022
	(0.112)	(0.153)	(0.098)	(0.089)	(0.104)
Florida	-0.189*	0.231***	0.051	0.077	0.046
	(0.114)	(0.069)	(0.083)	(0.063)	(0.062)
Georgia	0.302***	0.283***	0.215***	0.107*	0.034
	(0.079)	(0.06)	(0.069)	(0.056)	(0.05)
Idaho	-0.351	-0.259*	-0.043	-0.241**	0.094
	(0.318)	(0.142)	(0.12)	(0.096)	(0.08)
Illinois	0.385***	0.157***	0.233***	-0.025	-0.069**
	(0.054)	(0.04)	(0.037)	(0.038)	(0.031)
Indiana	-0.018	0.113**	0.120**	-0.081	0.024
	(0.075)	(0.045)	(0.053)	(0.05)	(0.041)
Iowa	0.206**	0.281***	0.331***	0.07	0.075
	(0.105)	(0.058)	(0.08)	(0.05)	(0.056)
Kansas	-0.373***	0.005	0.097	0.06	0.008
	(0.126)	(0.079)	(0.082)	(0.072)	(0.058)
Kentucky	0.048	0.302***	0.195**	0.032	0.01
	(0.109)	(0.109)	(0.085)	(0.043)	(0.064)

Note: This table presents a sample of the state coefficients from eqn. (4) for each year using only state level data. For ease of presentation standard errors are omitted, but stars are included to indicate the usual 90, 95, and 99 percent t statistic thresholds. R^2 was over 0.95 for all regressions and was suppressed for ease of presentation. All data come from the US Census of Manufactures. Data on Oregon were not scanned for the US 1909 Census of Manufactures.

Because of the size of the table, the full results from equation (4) using state level data are included in Appendix Table C. Table 4 shows a sample of the state fixed effects from equation (4) and their level of significance for each state in each year. This sample is sufficient to highlight to essential trends in the evolution of these state level skill premia over time.

The first trend is that there was a decline in the number of states with coefficients that were

statistically different from the national average. In this sample half of the states were statistically different from the national average at a 90% confidence level in 1900 and 1920. All but two were statistically different in 1910. However, by 1930 that number falls to only three states and by 1940 there is only one state that is statistically distinct from the national average. Though it is more difficult to glean from the table, the magnitude of the state level fixed effects is also generally, but evenly declining over time. These are signs of declining labor market frictions and evidence of an emerging homogeneity across state labor markets.

From Appendix Table C it is possible to identify the states that had persistently statistically distinct deviations from the national skill premium and what it meant for the development of industry in those states. Only two states, Missouri and New Hampshire, had skill premia that were distinct from the national average for the first four years. Missouri's skill premia was consistently larger than the national average meaning that firms in Missouri invested more in salaried work rather than wage work even after controlling for the distribution of Missouri's industrial composition. Firms in places like New Hampshire or Maine persistently supplied wage earners with higher relative payments than salaried workers.

By making assumptions about the constant returns to scale function for production it is possible to decompose these state level deviations into wage and employment components, however this is beyond the scope of this paper. Instead it is sufficient to highlight the fact that there were states with persistent differences from the national average. While there is some geographic consistency in the states with skill premia persistently below the national average from 1900 to 1930, there is not any amongst the seven states with skill premia persistently larger than the national average. Between 1900 and 1920 firms in Georgia, Illinois, Missouri, New York, Ohio, Tennessee, and Wisconsin pursued a strategy of paying salaried workers relatively more than wage earners compared to the national average.

While five of these states form a contiguous whole it is hard to see how these seven states represent any form of separated or poorly integrated labor markets. In particular, when examining only states with statistically significant deviations from the national average, there is no evidence

in any year of a Southern labor market that appears separate from the rest of the nation. This finding differs from existing narratives in the economic history that a national labor market did not emerge until the 1920s or 1930s and that in particular the Southern labor market was separate from the rest of the country.

Instead of finding geographic differences over time, visual inspection of Table 4 and Table 5 confirm that overall labor market integration did increase over time although unevenly. Table 5 counts the number of coefficients statistically different from the national average. It shows that between 1900 and 1920 there was effectively no change in labor market frictions, or no increase in labor market integration. Between 1920 and 1930 several states that had skill premia statistically different from the national average, and in particularly statistically larger than the national average, become statistically indistinguishable from the national average. What Table 5 also highlights is the abrupt nature of the change. These results show that labor market integration in the US was not slow and gradual in the 20th century, but was instead an abrupt transition between two potential equilibrium states.

Table 5: State Coefficients By Statistical Significance

Year	1900	1910	1920	1930	1940
Significant, Positive	10	11	10	3	2
Insignificant	35	35	35	43	44
Significant, Negative	4	3	4	3	3

Note: This table shows the number of α_r from eqn. (4) run on state level data that are statistically significant and in what direction they are statistically different.

We can also measure labor market integration by studying the levels of each coefficient instead of focusing solely on statistical significance. Figure 4 presents a kernel density plot of the coefficients from equation (4) for the state level data. Since the coefficients are constrained to sum to zero an economy with a more well integrated labor market is one with a distribution with more mass around zero. In the figure we still see what appear to be two distinct labor market equilibria with the first three years having similar distributions with less mass around zero and then the last two years having more mass around zero.

A Kolmogorov-Smirnoff test of distributional equality also highlights how sharp the difference

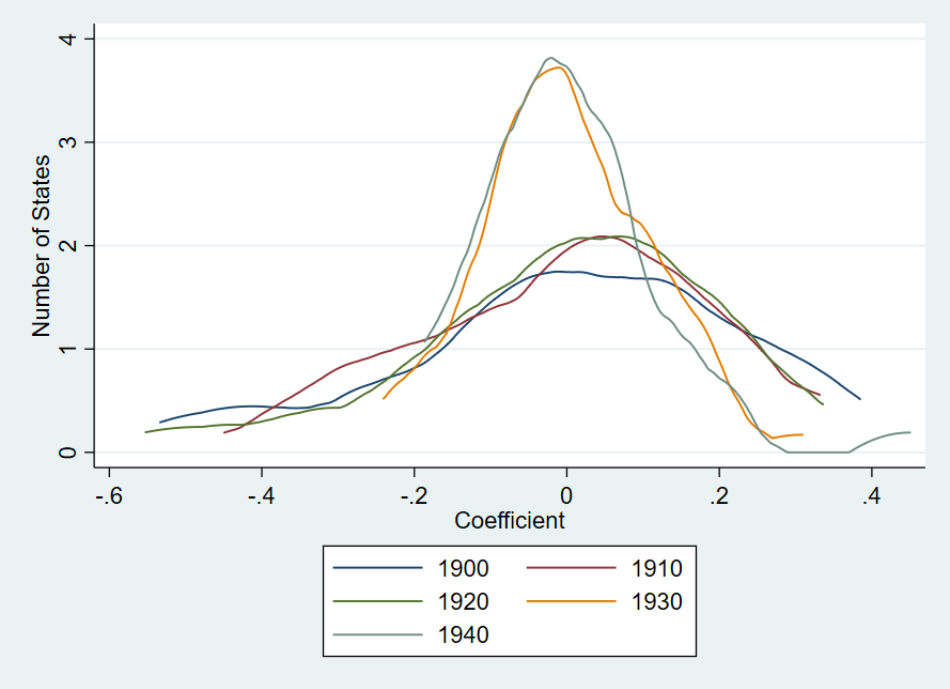


Figure 4: This figure presents the kernel density plot of state level coefficients from (4) for each year.

is between the distributions for 1900-1920 and 1930-1940. A summary of these tests is presented in Table 4. Each row of the table shows the p-value associated with rejecting the null hypothesis that the distribution of state fixed effects for each year are statistically different from each other.

To read Table 6 row by row we see that the distribution of state fixed effects for 1900 was much less likely to be the statistically different from the 1910 and 1920 distributions, than the 1930 and 1940 distributions. Restricting our attention to the 3x3 diagonal of entries in the top left of the table we see the lowest value is 0.65. The remaining entries show that the 1930 and 1940 distributions are much more likely to be statistically different from the 1900-1920 distributions and are nearly statistically identical to each other. All of this presents further evidence of a labor market over time with two distinct equilibria with the transition taking place between 1920 and 1930.

Finally, we can visualize the distribution of state fixed effects across space to identify any regions with apparently separate labor markets. Figure 5 presents the ordinal ranking of state fixed effects with darker colors reflecting states with skill premia larger than the national average and light colors reflecting states with skill premia below the national average.

Table 6: Kolmogorov-Smirnoff Test of Distributional Equality of State Fixed Effects

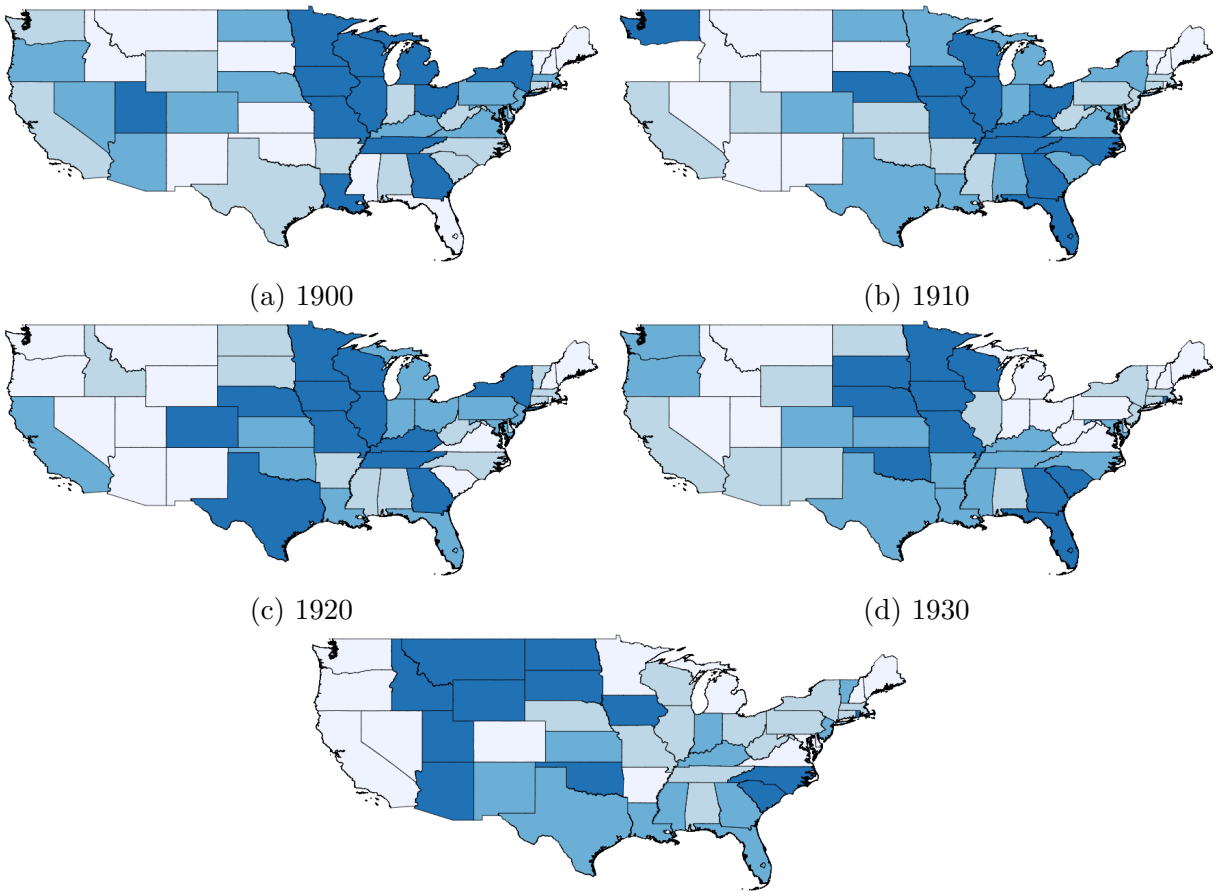
State	1900	1910	1920	1930	1940
1900	–	0.65	0.86	0.17	0.26
1910	–	–	0.9	0.3	0.13
1920	–	–	–	0.26	0.11
1930	–	–	–	–	0.99
1940	–	–	–	–	–

Note: This table presents the p values from a Kolmogorov-Smirnoff test associated with rejecting the equality of distributions of state level fixed effects eqn. (4) between each year. A value closer to one is associated with distributions that are less likely to be statistically different from one another.

Our formal test of labor market integration involves studying coefficients that are statistically distinct from the national average, but we may be able to identify meaningful regional trends by examining changes over time in the coefficients themselves. Unlike before where the greatest change took place between 1920 and 1930 in Figure 3 the most dramatic change takes place between 1930 and 1940. Until 1940 states in the Mountain census division had below national average skill premia. In 1940 these states have skill premia well above the national average. The sharp difference between the Mountain and Pacific census divisions may be worth investigating.

The only regional trend persistent across each year is that states in the Northeast persisently had below national average skill premia. Between 1900-1930 there was a corridor consisting of Minnesota, Wisconsin, Iowa, and Missouri that persistently had some of the highest skill premia, but in 1940 none of these states, but Iowa did. In none of the years do any of the Southern census divisions appear to be unique. Nor is there any time trend that shows states in any of the Southern census division becoming more integrated with the rest of the country.

Lastly, while the emphasis of this paper is that migration or lack thereof does not necessarily indicate the presence of labor market frictions scholars may desire to see how these measures of skill premia relate to other state level labor market data. Table 7 presents the results of a regression of skill premia on an array of state level controls that include the share of industrial employment in salaried work calculated using Census of Manufactures data. It also includes data from full count US censuses calculated by using all individuals with a known occupation, living in the United States, aged 16-65.



(a) 1900

(b) 1910

(c) 1920

(d) 1930

(e) 1940

Figure 5: These figures plot the state coefficients from eqn. (4) for each year. States are grouped into quartiles based on their coefficient. Darker colors indicate states with larger fixed effects.

Table 7 presents the results of a simple OLS regression of the share of a state's industrial workforce engaged in salaried work measured by the Census of Manufactures against the state level measure of skill premium equation (4) using state data and state level controls. For each state the controls include the share of the workforce that has their race recorded as white in the census, the share male, the share living in an urban area, the size of the labor force, and the share of labor force that was born outside the United States.

Table 7: Relationship Between Skill Premium and Salaried Share of Workforce

Variable	Share Salaried				
Year	1900	1910	1920	1930	1940
Ratio	-0.0004	-0.001**	-3e-05	0.001*	0.002**
	(0.0002)	(0.0005)	(0.0006)	(0.0007)	(0.0008)
White	-0.02***	-0.031***	-0.05***	-0.009*	-0.04***
	(0.003)	(0.005)	(0.007)	(0.005)	(0.011)
Male	0.01***	0.044***	0.077***	0.02***	0.12***
	(0.002)	(0.003)	(0.005)	(0.005)	(0.005)
Urban	0.046***	0.079***	0.127***	0.126***	0.215***
	(0.0037)	(0.006)	(0.00722)	(0.007)	(0.01)
Labor Force	-5e-08***	-2e-08***	-4e-08***	4e-08***	-4e-08***
	(6.e-09)	(4e-09)	(5e-09)	(6e-09)	(4e-09)
Immigrant	-0.01***	0.001	0.05***	0.02***	0.005
	(0.002)	(0.002)	(0.003)	(0.004)	(0.006)
Constant	-0.001	-0.042***	-0.07***	-0.04***	-0.09***
	(0.002)	(0.003)	(0.004)	(0.005)	(0.006)
Observations	2,579	2,046	2,502	2,378	2,469
R-squared	0.266	0.413	0.463	0.507	0.663

Note: This table presents results from an OLS regression of the share of state's labor force engaged in salaried work on the wage bill ratio of each state industry group pair. The regressions are run separately for each year.

These shares were computed using full count census data where the sample was restricted to all individuals in a few categories of industrial work and in relevant industries. Restrictions were necessary since census recordings of occupation and industry do not match the Census of Manufactures. While the results are weak there is some statistical evidence that states with a higher skill premium have more salaried workers. This effect becomes positive and weakly significant between

1920 and 1930, which matches the timing of the observed decrease in labor market frictions. [2]

5.1 Robustness Checks

This section includes several robustness checks to demonstrate that these findings are not an artifact of geography or industrial classification. First, kernel density plots for city level and state and city level data produce similar results that labor market integration increased between 1920 and 1930. Next, the kernel density plot for 1920 using state level and establishment level data show that a finer level of industrial activity does not change the distribution of state level coefficients. Finally, the kernel density plot of state level coefficients using a two digit SIC classification system shows that the results are not an artifact of overly fine geography.

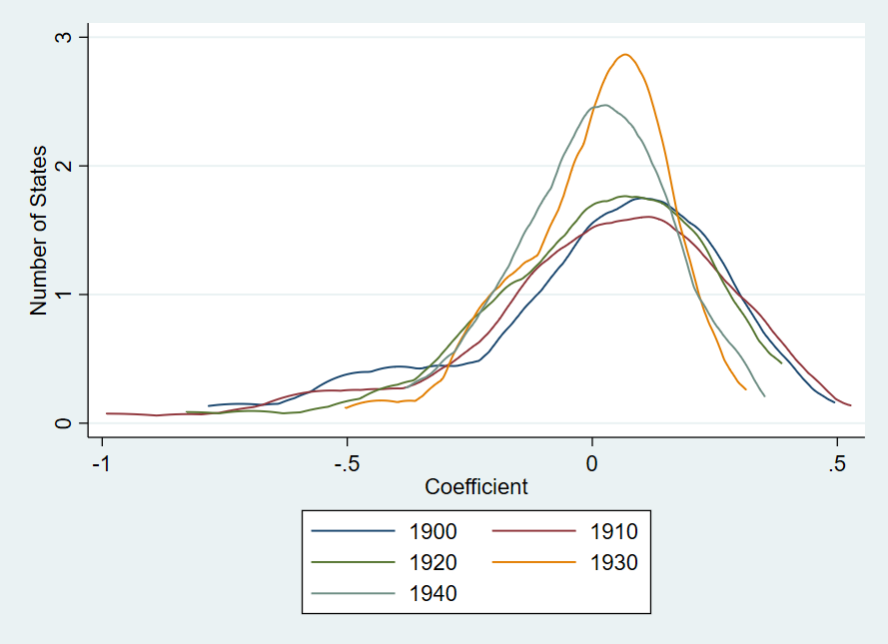
Figure 6 shows that whether the level of geography is all US states, a panel of 44 cities, or a combination of the two where states have city values removed to avoid double counting there is similar behavior in the region fixed effects over time. In all figures the distributions for 1930 and 1940 are similar and visually distinct from the distributions for 1900, 1910, and 1920.

Instead of relying on industry level data it is also possible to use establishment level data to estimate equation (4). The advantage of using establishment level data is that it allows for variation within each industry and state rather than relying on variation of industries across states. The disadvantages of using establishment level data are that it is only for some industries, is not available for all states, and is not available in all years.

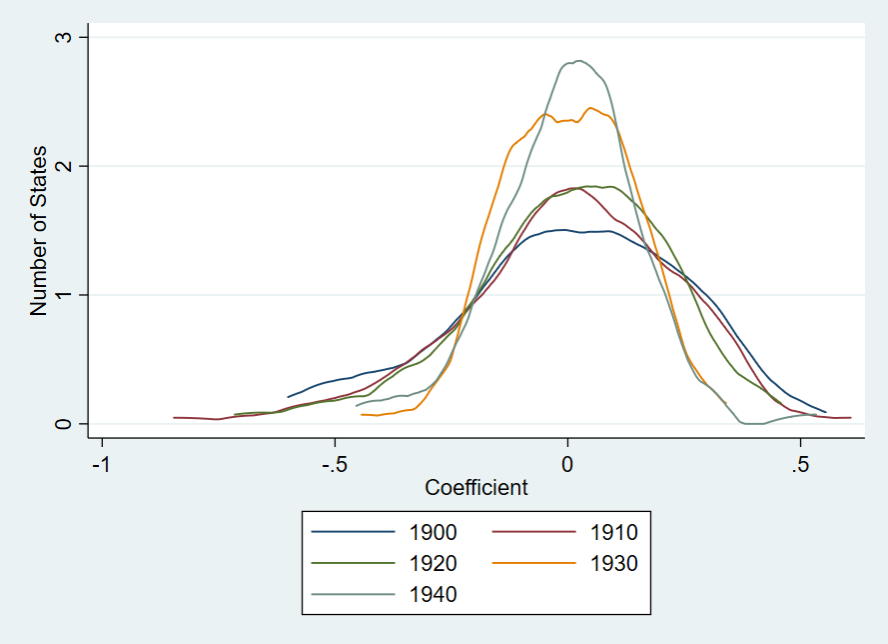
Because of these disadvantages establishment level data is only used to compare state coefficients in 1930. The establishment level data include fewer industry groups and states than the primary data. To harmonize the geography only states included in the establishment level data were used in the primary state level analysis. Because substantially fewer industry groups are covered by the establishment data no effort was made to harmonize the industry groups between the two sets.

For the establishment level data equation equation (4) is estimated using the full suite of industry and state fixed effects. The only difference between the two regressions is that the establishment

[2] Workers were restricted to those with occ1950 = 390, 690, 970 with 390 being salaried workers. Workers were restricted to industries with $299 < \text{ind1950} < 500$.



(a) Cities



(b) Cities and States

Figure 6: This figure shows the kernel density plots for the city level and state plus city level data estimated from eqn. (4) described in the data section of the paper.

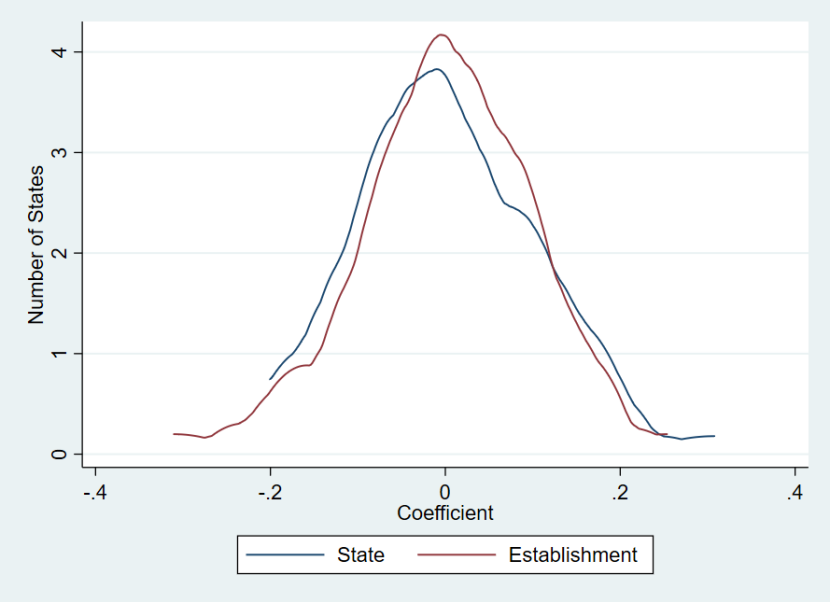


Figure 7: This figure shows the kernel density plots of 1930 state level coefficients estimated using industry level data against state level coefficients estimated using establishment level data. In this setting Idaho and D.C. are excluded since they are excluded in the establishment level data. Establishment level data come from ICPSR.

level data have more entries per state-industry pair. Per Figure 7, the distribution of state fixed effects have not significantly changed even with this finer level of industrial detail. A Kolmogorov-Smirnoff test of distributional equality says that the two distributions are as identical as the 1930 and 1940 distributions from state level data.

The final robustness check is the state level analysis done with a coarser level of industrial grouping. Figure 8 is the kernel density plot of state fixed effects estimated at the two digit SIC level. The basic trend highlighted in the other kernel density plots is present here. The distributions for 1930 and 1940 appear similar, but distinct from the distributions for 1900-1920 which are roughly similar. While Figure 6 and Figure 4 highlight the same trends, they also highlight the importance of having a high level of industrial variation for this estimation strategy. Again, A Kolmogorov-Smirnoff test yields similar, albeit weaker results, compared to the standard state coefficients.

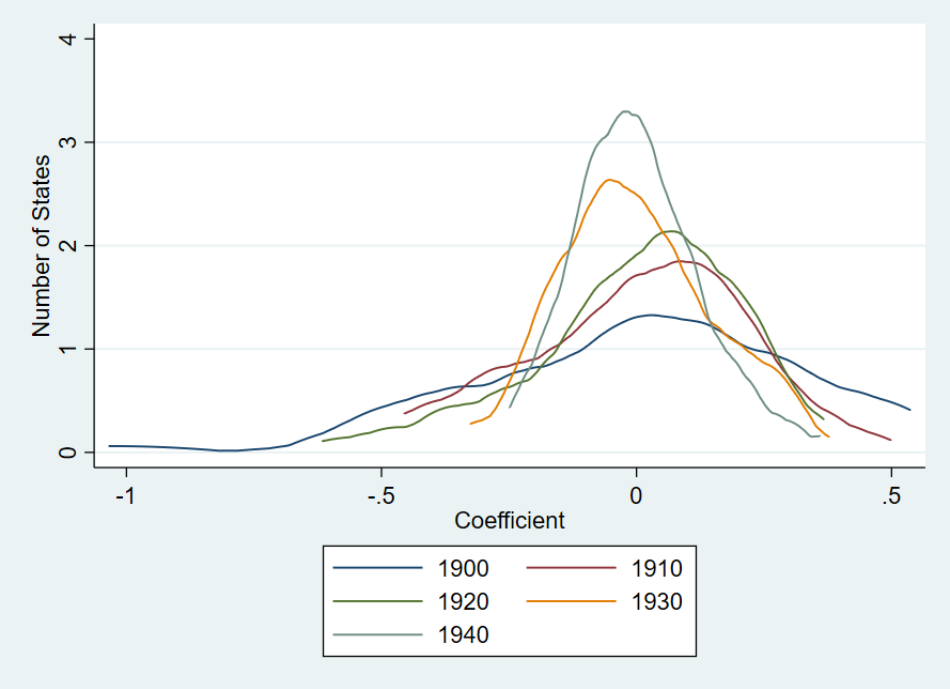


Figure 8: This figure presents the kernel density plot of state fixed effects from estimating eqn. (4) at the two digit SIC level rather than the three digit SIC level.

6 Discussion

This paper uses new and more comprehensive data along with a more robust measure of labor market integration to show that labor market integration increased between 1900 and 1940 and in particular between 1920 and 1930. Finding increased labor market integration between 1900 and 1940 is consistent with the argument advanced by Joshua Rosenbloom who argues that labor market integration followed in the wake up of the border closures of the early 1920s that caused a significant decrease in the labor supply for Northern industry. Rather than rely on raising wages, Northern firms began to recruit low wage Southern labor which served as the prelude for the Great Migration that led to labor market integration between the North and South. In this version of events Northern hesitancy against using Southern labor was a major force that inhibited labor market integration.

However this paper finds no evidence that labor markets in Southern states were different from labor markets in other parts of the country in a way that imply their labor markets were disaggre-

gated from the rest of the country. The firms in states that focused on attracting salaried workers above and beyond the national trend were geographically diverse. Instead this paper presents weak evidence that firms in Western states either tended to offer particularly low relative payments to salaried workers or relatively high payments depending on the year.

The two most likely reasons for this difference in findings are both data related. The first is the fact that existing estimates of real wages relied on incomplete data and incomplete estimates of price indexes. Even if these data were complete, there are reasons to be skeptical of drawing conclusions about labor markets by only looking at industrial wage earners. One is related to the unobserved productivity terms being ignored by studies of wages that relying on wage bills solves. The second is that industrial wage earners are only one part of the labor market and even only one part of the industrial labor market.

Additionally, the effort to highlight and understand the fact that Southern labor markets were different from the rest of the country is laudable, but incorrectly labelling these differences as a sign of poor labor market integration clouds our ability to properly interpret these differences. For example, no one would look at the difference in crops grown across the United States and accuse agricultural markets of being poorly integrated - agricultural products are dependent on region specific qualities such as climate and soil quality. If economic historians can tolerate these kinds of differences it seems odd that the existence of labor market differences would be so vexing. Furthermore, there is no literature accusing western labor markets of being poorly integrated with the North even though the kinds of migration needed to equalize wages also took time to materialize.

Part of the reason why economic historians have looked for labor market disintegration is born out of a desire to explain the Great Migration and in particular why it happened when it happened. This is an open question and given the importance of the Great Migration in American history an important one. But, we know that an absence of migration is not necessarily evidence of a lack of labor market integration. Based on existing estimates of real wages, Western states had persistently high real wages over this period. Whatever westward migration took place during this period was not sufficient to equalize Western and Eastern wages. Where are the significant enough

labor market frictions between the West and East to qualify the West as a separate labor market?

Based on the findings in this paper we can learn more about national labor market integration by focusing on national, rather than regional trends. We know that between 1920 and 1930 firms across the United States began to value salaried work relative to wage work in a much more similar fashion. Unfortunately, because the estimated skill premia were not geographically correlated it is not possible to draw the same kind of regional conclusions previous scholars did about wages. However, it may be that earlier wage data were incomplete enough to hide the substantial amount of intra-divisional variation. My hope is that future work on emerging national forces in labor market integration can provide additional clarity on this topic. If scholars can find ways to use this new data to do so, all the better.

7 Works Cited

Abramitzky, Ran, and Leah Boustan. "Immigration in American Economic History." Journal of Economic Literature, vol. 55, no. 4, 2017, pp. 1311-1345.

Abramitzky, Ran, Philipp Ager, Leah Boustan, Elior Cohen, Casper Worm Hansen. "The Effects of Immigration on the Economy: Lessons from the 1920s Border Closure." National Bureau of Economic Research. 2019.

Anderson, James E., and Eric van Wincoop. "Gravity with Gravitas: A Solution to the Border Puzzle." The American Economic Review, vol. 93, no. 1, 2003, pp. 170-192.

Bernard, Andrew B., Stephen J. Redding, and Peter K. Schott. "Testing for Factor Price Equality with Unobserved Differences in Factor Quality Or Productivity." American Economic Journal. Microeconomics, vol. 5, no. 2, 2013, pp. 135-163.

Collins, W., & Wanamaker, M. (2015). The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants. The Journal of Economic History, 75(4), 947-992.

Collins, Williams J. and Ariell Zimran. "Immigrants' Changing Labor Market Assimilation in the United States during the Age of Mass Migration." National Bureau of Economic Research 2019.

Collins, William J. "The Great Migration of Black Americans from the US South: A Guide and Interpretation." Explorations in Economic History, vol. 80, 2021, pp. 101382.

Ferrie, Joseph P. "History Lessons: The End of American Exceptionalism? Mobility in the United States since 1850." The Journal of Economic Perspectives, vol. 19, no. 3, 2005, pp. 199-215.

Friedman, Milton, and Anna J. Schwartz. A Monetary History of the United States, 1867-1960. vol. no. 12.;12.;, Princeton University Press, Princeton, N.J, 1971.

Kennedy, David M. Freedom from Fear: The American People in Depression and War, 1929-1945. vol. 9., Oxford University Press, New York, 1999.

Kim, Sukkoo. Economic Integration and Convergence: U.S. Regions, 1840-1987. The Journal of Economic History, Vol. 58, No. 3, 1998, pp. 659-683.

McCallum, J. (Royal Bank of Canada, Toronto, Ontario, Canada.). "National Borders Matter: Canada: U.S. Regional Trade Patterns." The American Economic Review, vol. 85, no. 3, 1995, pp. 615-623.

Mitchener, Kris J., and Ian W. McLean. "U.S. Regional Growth and Convergence, 1880-1980." The Journal of Economic History, vol. 59, no. 4, 1999, pp. 1016-1042.

Ran Abramitzky, Leah Boustan and Myera Rashid. Census Linking Project: Version 1.0 [dataset]. 2020. https://censuslinkingproject.org

Rosenbloom, Joshua L. Looking for Work, Searching for Workers: American Labor Markets during Industrialization. Cambridge University Press, Cambridge, UK;New York;, 2002.

Rosenbloom, Joshua L. "One Market Or Many? Labor Market Integration in the Late Nineteenth-Century United States." The Journal of Economic History, vol. 50, no. 1, 1990, pp. 85-107.

Salisbury, Laura. Selective migration, wages, and occupational mobility in nineteenth century America. Explorations in Economic History. 2014.

Steinwender , Claudia. "Real Effects of Information Frictions: When the States and the Kingdom Became United." American Economic Review, vol. 108, no. 3, 2018, pp. 657-696.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 9.0 [dataset]. Minneapolis, MN: IPUMS, 2019. https://doi.org /10.18128 /D010.V9.0

Wright, Gavin. Old South, New South: Revolutions in the Southern Economy since the Civil War. Basic Books, New York, 1986.

Wright, Gavin. "The Economic Revolution in the American South." The Journal of Economic Perspectives, vol. 1, no. 1, 1987, pp. 161-178.

8 Appendix

Appendix Table A: Cities

State	City	State	City
AL	Birmingham	NJ	Jersey City
CA	Los Angeles	NJ	Newark
CA	Oakland	NJ	Paterson
CA	San Francisco	NJ	Trenton
CO	Denver	NY	Buffalo
CT	Bridgeport	NY	New York
CT	New Haven	NY	Rochester
DE	Wilmington	NY	Syracuse
IA	Des Moines	NY	Utica
IL	Chicago	OH	Cincinnati
IN	Evansville	OH	Cleveland
IN	Indianapolis	OH	Columbus
LA	New Orleans	OR	Portland
MA	Boston	PA	Philadelphia
MA	Cambridge	PA	Pittsburgh
MA	Fall River	PA	Reading
MA	Lowell	PA	Scranton
MD	Baltimore	TX	Houston
MI	Detroit	TX	San Antonio
MI	Grand Rapids	WA	Seattle
MN	St Paul	WA	Spokane
MO	St Louis	WA	Tacoma

Note: This table lists the panel of 44 cities used in my analysis.

Appendix Table B: Industry Groups

ID	Industry Group
201	Meat
202	Dairy
203	Canning & Preserving Food
204	Grain Preparation
205	Bread & Bakery Production
206	Sugar, Gum, Chocolate, and Nut Preparation
207	Oils and Animal Fats
208	Alcohol, Flavoring Syrups, and Other Beverages
209	Misc Food
223	Dyed Fabric, Rayon, Silk, Wool, Underwear
224	Cotton Goods, Handkerchiefs, and Blinds
225	Knit, Lace, Hosiery, and Curtains
226	Cloth, Sponging
227	Carpets, Rag, Fibrous Products
229	Waste, Artificial Leather, Felt, Flooring
232	Men's Clothing
233	Women and Children's Clothing
234	Corsets
235	Hats and Millinery
237	Furs
238	Gloves and Mittens, Regalia, Trousers, MISC apparel
239	Other Clothing
242	Logging Products
243	Window Shades and Door Screens
244	Wooden Boxes and Cooperage
249	Other Wood Products
251	Mattresses and Beds
252	Furniture, Upholstery, Refrigerators
261	Wood Pulp
262	Paper Products
265	Paper Boxes
267	Bags, Stationery Goods, Wallpaper, Files
279	Stereotyping and Electrotyping, Bookbinding, Cards
281	Salt, Acids, and Compressed Gasses
283	Medicine
284	Cleaning Products, Perfumes, Blacking
285	Paint, Varnish, Signs
286	Charcoal, Turpentine, Natural Dyestuffs
287	Fertilizer and Insecticide
289	Miscellaneous Chemical Products
291	Petroleum, Paving, Roofing
299	Coke, Grease, Lubricating Oils
301	Rubber Products Other than Boots
302	Rubber Boots
305	Steam fitting, packing, and rubber belting
311	Leather
313	Boot and Shoe Cut Stock
319	Miscellaneous Leather Goods
323	Mirrors
324	Cement, Lime, Concrete
325	Clay, Stone, Brick
326	Pottery and Similar Products
328	Marble, Stone, Artificial Products
329	Miscellaneous Minerals
331	Wire
332	Pipe, Cast-Iron, Castings, Repair
342	Cutlery, Hardware, Tools
343	Stoves, Heating
345	Bolts, Screws, Forgings
347	Miscellaneous Ornamental Work
348	Firearms and Ammunition
352	Agricultural Implements (Except Tractors)
355	Miscellaneous Trade Machinery
356	Pumps, Washing Machines, Miscellaneous Machinery
359	Scales and Balances
363	Domestic Equipment
365	Photography
369	Miscellaneous Electronics
371	Automobile, Internal Combustion Engine, Windmill
373	Shipbuilding
374	Trains, Other Transportation
375	Motorcycles and Bicycles
382	Measuring Instruments
384	Professional Supplies
387	Clocks and Watches
391	Jewelry, Cigar Boxes
393	Musical Instruments
394	Recreation Goods
395	Writing Supplies, Cash Registers
396	Needles, Buttons, Theatre
399	Other Goods
901	Tobacco Products
902	Foils
903	Safes and Vaults
904	Springs
905	Vinegar and Cider
906	Ice, Manufactured
907	Miscellaneous Metals
910	House Furnishings
915	Printing and Publishing and Matches
918	Chemicals
921	Glass
931	Lighting
936	Brooms, Brushes
939	Coffins
953	Models
954	Photo-engraving
959	Umbrellas, Canes
961	Ice Cream and Confectionery

Note: This table includes the full list of industrial groups between 1899-1939 inclusive used in this paper.

Appendix Table C: Primary Regression Results for US States

Variable			Ratio						Ratio		
Year	1900	1910	1920	1930	1940		1900	1910	1920	1930	1940
AL	0.015	0.111	0.018	-0.027	-0.047	NE	0.088	0.283***	0.134	0.309***	-0.031
	(0.094)	(0.099)	(0.079)	(0.076)	(0.076)		(0.117)	(0.092)	(0.082)	(0.08)	(0.069)
AZ	0.032	-0.187	-0.252	-0.03	0.176**	NV	0.041	-0.236	-0.466*	-0.149	-0.186
	(0.167)	(0.171)	(0.175)	(0.091)	(0.08)		(0.476)	(0.168)	(0.247)	(0.172)	(0.193)
AR	-0.031	-0.056	-0.018	-0.015	-0.095	NH	-0.508***	-0.449***	-0.311***	-0.199**	-0.104
	(0.131)	(0.119)	(0.121)	(0.084)	(0.064)		(0.116)	-0.099	(0.081)	(0.078)	(0.073)
CA	0.012	-0.055	0.049	-0.068*	-0.084**	NJ	0.173**	-0.017	0.05	-0.052	0.009
	(0.07)	(0.071)	(0.055)	(0.038)	(0.036)		(0.084)	(0.058)	(0.048)	(0.04)	(0.04)
CO	0.041	0.036	0.167**	0.053	-0.139*	NM	-0.115	-0.294	-0.552***	-0.047	0.04
	(0.106)	(0.087)	(0.074)	(0.064)	(0.078)		(0.223)	(0.189)	(0.151)	(0.114)	-0.137
CT	-0.009	0.001	-0.014	-0.031	-0.004	NY	0.264***	0.117***	0.163***	-0.017	-0.066**
	(0.062)	(0.055)	(0.065)	(0.046)	(0.048)		(0.047)	(0.045)	(0.037)	(0.03)	(0.033)
DE	-0.05	-0.086	-0.074	-0.05	-0.089	NC	-0.106	0.120	-0.026	0.012	0.07
	(0.142)	(0.115)	(0.092)	(0.112)	(0.077)		(0.123)	(0.091)	(0.07)	(0.067)	(0.05)
DC	-0.165	-0.374**	0.131	0.18**	0.022	ND	0.03	0.0337	0.007	-0.02	0.45***
	(0.112)	(0.153)	(0.098)	(0.089)	(0.104)		(0.248)	(0.180)	(0.162)	(0.06)	(0.077)
FL	-0.189*	0.231***	0.051	0.077	0.046	OH	0.242***	0.144***	0.098***	-0.068	-0.05
	(0.114)	(0.069)	(0.083)	(0.063)	(0.062)		(0.064)	(0.046)	(0.038)	(0.043)	(0.036)
GA	0.302***	0.283***	0.215***	0.107*	0.034	OK	-0.455**	-0.146	0.102	0.161**	0.209***
	(0.079)	(0.06)	(0.069)	(0.056)	(0.05)		(0.197)	(0.1)	(0.092)	(0.074)	(0.076)
ID	-0.351	-0.259*	-0.043	-0.241**	0.094	OR	0.163		-0.184**	-0.01	-0.186***
	(0.318)	(0.142)	(0.12)	(0.096)	(0.08)		(0.111)		(0.093)	(0.073)	(0.064)
IL	0.385***	0.157***	0.233***	-0.025	-0.069**	PA	0.107**	0.017	0.0883**	-0.106***	-0.05
	(0.054)	(0.04)	(0.037)	(0.038)	(0.031)		(0.054)	(0.048)	(0.039)	(0.031)	(0.033)
IN	-0.018	0.113**	0.120**	-0.081	0.024	RI	-0.144*	-0.174	-0.123	0.147**	0.106**
	(0.075)	(0.045)	(0.053)	(0.05)	(0.041)		(0.077)	(0.107)	(0.078)	(0.065)	(0.053)
IA	0.206**	0.281***	0.331***	0.07	0.075	SC	0.005	0.027	-0.131	0.120	0.120
	(0.105)	(0.058)	(0.08)	(0.05)	(0.056)		(0.087)	(0.114)	(0.112)	(0.101)	(0.087)
KS	-0.373***	0.005	0.097	0.06	0.008	SD	-0.533***	-0.251	-0.066	0.145	0.187
	(0.126)	(0.079)	(0.082)	(0.072)	(0.058)		(0.192)	(0.159)	(0.133)	(0.116)	(0.114)
KY	0.048	0.302***	0.195**	0.032	0.01	TN	0.262**	0.259***	0.19***	0.038	-0.054
	(0.109)	(0.109)	(0.085)	(0.043)	(0.064)		(0.107)	(0.075)	(0.058)	(0.041)	(0.046)
LA	0.3***	0.105	0.106	0.031	0.014	TX	-0.108	0.109	0.164***	0.063	0.043
	(0.08)	(0.094)	(0.076)	(0.067)	(0.053)		(0.111)	(0.073)	(0.061)	(0.05)	(0.038)
ME	-0.385***	-0.158*	-0.236***	-0.083	-0.151*	UT	0.195*	-0.009	-0.261**	-0.125*	0.171*
	(0.130)	(0.083)	(0.072)	(0.064)	(0.08)		(0.115)	(0.091)	(0.109)	(0.071)	(0.096)
MD	0.081	0.078	0.052	0.019	-0.113	VT	-0.141	-0.207***	-0.091	-0.182*	0.037
	(0.07)	(0.064)	(0.059)	(0.049)	(0.073)		(0.096)	(0.076)	(0.056)	(0.101)	(0.081)
MA	0.069	-0.079*	-0.0216	-0.024	-0.014	VA	0.133	0.0422	-0.130*	-0.068	-0.071
	(0.064)	(0.048)	(0.04)	(0.037)	(0.041)		(0.104)	(0.077)	(0.074)	(0.056)	(0.054)
MI	0.246***	0.066	0.095*	-0.076	-0.077*	WA	-0.013	0.128*	-0.112	0.041	-0.179***
	(0.07)	(0.054)	(0.05)	(0.052)	(0.043)		(0.099)	(0.067)	(0.075)	(0.051)	(0.056)
MN	0.236***	0.105*	0.336***	0.124**	-0.176***	WV	-0.115	0.005	-0.015	-0.189**	-0.038
	(0.08)	(0.064)	(0.055)	(0.056)	(0.062)		(0.117)	(0.063)	(0.093)	(0.08)	(0.064)
MS	-0.156	-0.038	-0.066	0.036	0.03	WI	0.197***	0.120**	0.214***	0.153***	-0.046
	(0.12)	(0.106)	(0.071)	(0.063)	(0.081)		(0.065)	(0.055)	(0.048)	(0.055)	(0.041)
MO	0.373***	0.332***	0.329***	0.179***	-0.029	WY	0.003	-0.254	-0.422***	-0.063	0.105
	(0.069)	(0.055)	(0.054)	(0.039)	(0.038)		(0.228)	(0.186)	(0.134)	(0.117)	(0.152)
MT	-0.282*	-0.285	-0.122	-0.112	0.065						
	(0.158)	(0.218)	(0.101)	(0.091)	(0.066)						
Obs	2,579	2,046	2,502	2,378	2,469	Obs	2,579	2,046	2,502	2,378	2,469

Note: This table presents the full results from eqn. (4) using state level data from the US Census of Manufactures.

8.1 Data Construction

This appendix details how the new census of manufactures data used in this paper were constructed. To build the full set of 23,404 state level and 12,931 city level entries of industrial activity publicly available scans of the US Census of Manufactures for each decade between 1900-1940 inclusive were used. These scanned reports are available by state or by industry. Contained within the reports by state for the years 1910-1940 are information on the number of establishments, the number of salaried/wage earners, the wage bills for both groups, the value of the industry, and the value added of the industry. In 1900, all of those variables except value added were recorded since value added was not included in 1900. For most years enough variables were included that each line of industrial activity was present on two pages. In this appendix I refer to a first and second page when discussing which variables from the Census of Manufactures are used to construct the data. I also refer to listed variables which are from the Census and recorded variables which are what were recorded or constructed from the data. [3]

Even though the same variables were recorded for each year, the data listed by the Census and the variables used to create the recorded data changed year to year. The 1900 data came from the *Manufactures by Specified Industries and Manufactures in Cities* tables at the end of each state's chapter of the Census. All relevant variables, except for the variable value, were taken from the first page of the table. For establishments, the listed number of establishments was used. For the number of salaried workers the listed number of salaried officials, clerks, etc was used. Proprietors or firm members were not included. For the number of wage earners the listed average was used. For the relevant wage bills listed salaries and wages were used. Value was taken from the second page of the table. The 1900 tables contain information that were not collected regarding capital and other expenditures as well as a break down by gender and youth for wage employment. In total 5,270 entries for the continental states and DC and 3,741 entries for the panel of 44 cities were recorded for 1900.

[3] I hand recorded data for all of the continental states and DC for each year and the city level data for 1940. I thank Digital Divide Data for their help in recording the city level data for 1900-1930 inclusive.

The 1910 data come from the *Detailed Statement For the State, By Industries* as well as the *Detailed Statement for Cities*. Like in 1900, the table consists of a front and back with the front listing the number of establishments and levels of employment, while the back includes wage bill and value data. For establishments the listed number was used. For salary workers the fields for salaried officers, male workers, and female workers were summed. Proprietors and firm members were not used. For wage earners the listed average was used. The salary wage bill was the sum of the wage bill for officials and clerks. The wage earner wage bill was the listed number. The listed numbers for value and value added were used. The 1910 data included many of the same variables as the 1900 data. In total 2,725 entries for the continental states and DC and 1,280 entries for the panel of 44 cities were recorded for 1910.

The tables for 1920 were similar to 1910. The names of the tables as well as the listed variables used to construct the recorded variables are identical between 1910 and 1920. The differences between the tables are in some of the unrecorded variables, such as horsepower devoted to each industry. The other difference is the inclusion of sub-categories for industrial activity. For example, in other years men's clothing would be broken down into several different entries on the table without any being given precedence over the other. The 1920 tables included a main heading for men's clothing, sub-headings for different categories of men's clothing, and then the individual entries. To avoid double counting only the lowest level of industrial activity for any group was recorded. In total, 4,670 entries for the continental states and DC and 2,565 entries for the panel of 44 cities were recorded for 1920.

The tables in the 1930 data deviate from the previous years because there are fewer listed variables such that each entry fits on only one page. The name of the table changed to *General Statistics For the State, By Industries or General Statistics For Cities*. The listed number of establishments, salaried officers and employees, wage earners (average) as well as the listed salaries, wages, value of products, and value added were recorded directly from the Census. The additional variables included in this Census are costs of materials, costs of fuels, prime movers, and a horsepower breakdown between prime movers and electric motors. The limited scope of the 1930 data are the main

reason additional variables were not included in my data. In total 4,194 entries for the continental states and DC and 1,824 entries for the panel of 44 cities were recorded for 1930.

The tables in the 1940 data keep the same name as the 1930 tables, but include more variables for the state level data. Information on the number of establishments and employment are on the first page and information on the wage bills, value, and value added are on the second page. The same approach used for the 1930 data was used for the 1940 when matching listed to recorded variables. The one difference is that in 1940 the Census records the number of salaried officers of corporations and their wage bill. These workers were not included in either the recorded count of salaried workers or the salaried worker wage bill. For cities the 1940 data include fewer variables such that all of them fit on one page. The same convention for matching Census fields to the data was used here as it was for the states. In total 6,545 entries for the continental states and DC and 3,521 entries for the panel of 44 cities were recorded for 1940.

Each year the Census included different cities from each state to report. Researchers may find value in recording data from all of the cities in all years, but my research design demanded geographic consistency. The panel of 44 cities were determined entirely by which cities were available in all five years. There are tables for Bridgeport, Connecticut in all years, but in the 1900 Census one of the pages is missing for half the entries and thus Bridgeport was excluded completely from the analysis. The last page is missing for New York City in 1920, but New York City was so important in each of the years and so much was still recorded for 1920 that it was not excluded from any year. Whether an issue in the original document or the scanning process the pages for the state of Oregon are not present for the 1910 Census. The decision to only record some variables from the Census was a function of resource constraints, not a mark against the quality of the other data.

8.2 Assigning Industry Groups

For each year and state the creators of the Census had to determine what industrial activity was relevant enough for inclusion with many smaller groups being relegated to a table footnote with only the number of establishments listed. This approach meant it was not possible to create a

separate industry group for each line of industrial activity since the names of the same activity not only changed between years, but were sometimes recorded differently for cities and states. In particular, because my research design requires a stable panel of industry groups through time, the goal in creating industry groups was to ensure no activity was excluded while keeping groups as granular as possible.

The first phase of the project began by recording the 1920 state level data and then using an OSHA online web search to identify which two, three, and four digit SIC level each kind of industrial activity belongs to. Because so much of the recorded activity did not fit a four digit SIC category the three digit level was selected for primary use. In cases where industry activity could not be assigned a three digit code a placeholder code was used to best group similar activity. The best example is how the production of different kinds of tobacco products were recorded by some states that separated cigars and cigarettes and other states that reported them together. In this case a placeholder group was used that included all tobacco products. The full list of industry groups is included in Appendix Table B.

After recording all of the 1930 state level data the process of searching for and assigning industrial activity to industry groups based on the OSHA three digit classification system was repeated. As expected, the industry groups present in both years were different due to evolution in the industrial economy and changes in naming conventions. To harmonize a set of time consistent industry groups between the two years groups that were present in 1920, but missing in 1930 and present in 1930, but missing in 1920 were identified. Next either a placeholder group was created to group the data or a similar enough three digit code was used. Because of this, even the industry groups with a three digit SIC code may include industrial activity not found in the standard OSHA classification.

This process was repeated as the 1940, 1910, and 1900 data were completed. Each time a new year was transcribed a new master set of all previous years was constructed until there was a set of industry groups present in at least one state in every year. Each time a new set of data was added the primary analysis was repeated. The distribution of state fixed effects were robust to these changes in industry groups. There was attrition in the total number of industry groups as

new years were added from an initial 130 to the final 99 groups. To clarify, while each of the 99 industry groups is present in each of the years, each industry group is not present in each state so there is not a balanced panel in industry groups across time.

City level data were recorded once all state level data had already been recorded. Because the listed state level industrial activity should cover any of the listed data for cities, the focus was on matching listed industrial activity to listed state industrial activity and then using the industry group assigned to the industrial activity for the state. In instances where cities recorded industrial activity that was included in the other category at the state level the OSHA web search was used to assign a code and if that failed, the activity was assigned to the most similar industry group. Using this approach meant that not all industry groups are present in at least one city for every year. In practice because industrial activity was concentrated in cities nearly all industry groups are present in all years. While the city data are of similar quality to the state data researchers should exercise caution when using city level data since the boundaries of cities changed more frequently than the boundaries of states did in the early 20th century.

The combined set of all states and the panel of 44 cities was constructed by first taking the set of states and subtracting out all industrial activity attributable to one of the cities in the panel. Then this was combined with the existing city level data. The combined data set maintains geographic consistency, avoids double counting city level activity, and nearly doubles the number of geographic regions included in the analysis. The drawback of the combined set is that there are instances where industrial activity was recorded at the city level, but included in the other category at the state level. Thus there are cases where performing the above analysis leaves states with negative values for some industry groups. This is most common in the 1940 data because of the way it listed the other category. In all cases, negative values were zeroed out after being reviewed.

8.3 Data Verification and Data Quality Concerns

In a project like this with 36,335 rows of industrial activity and eight variables recorded by hand there is the potential for transcription error. Transcription error is important to minimize because

it can lead to biased estimates of regressors if the probability of error is correlated with one of the regressors. Even if the transcription error is uncorrelated with any regressors it still leads to attenuation bias that causes standard errors to be larger. This section explores the sources of transcription error, the implication for analysis, and issues in identifying transcription error in the 1900 and 1940 data.

There are two potential sources for transcription error. The first are transcription errors in the documents themselves which will be referred to as listed transcription error. These errors arise either from mistakes in the original transcription by Census enumerators or errors associated with scanning the documents. Examples of these errors include listed value added fields being larger than corresponding listed value fields, ink splotches or white spaces that cover up listed values, or the exclusion of data on Oregon from the 1910 Census. Even when there was reason to believe the listed values were wrong they were kept to best match the listed values in the Census.

It is not possible to identify all of the instances of listed transcription error. However, listed transcription error is likely to not be correlated with any regressors of interest. Outside of the 1900 data discussed below, the quality of scanned documents got better over time, but not significantly. Cases of stains or white marks were rare and did not appear concentrated in any one city or state. Transcription errors may have been more common in larger states or cities, but that is purely because the larger the number an enumerator records, the more opportunities there are to record an incorrect number.

The second kind of transcription error is what will be referred to as recorded transcription error. These arise when a number is recorded that does not match the listed value in the Census. Three techniques were employed to minimize the opportunity for these errors. The first technique was how the data were recorded. Data were entered column by column for all years except city level data for 1940. Working column by column ensured that data could be easily reviewed as it was entered and ensured more rows could not be recorded for one variable over another. It also helped prevent skipping values or recording values from an adjacent column.

The next two techniques were performed after all data had been recorded. The second technique

was an outlier test. For each year and each variable the largest and smallest values were reviewed until five values in a row were confirmed. All instances where value added was smaller than value and where the sum of salaries and wages exceeded value were also investigated. The final technique involved comparing the listed sums for each variable by state and year to the sums of my recorded values for each variable by state and year.

Checking the sums for these seven variables for all of the cities and all of the states in every year yielded 3,155 total fields that needed to be reviewed. Nearly half of the fields had no difference between my recorded value and the sum listed by the Census. Around 1,250 entries had less than a 1% difference. The remaining 400 entries had more than a 1% difference. Every instance of a difference of 1% or greater was reviewed and when appropriate a correction was made or the existing values were confirmed. The document TotalCheck, which can be found in the online appendix, documents this process. Cells with differences greater than 1% and were deemed valid are highlighted green. Cells with a greater than 1% difference that cannot be verified are highlighted in yellow.

There are features of the data that vary by year and make it more likely for transcription error to take place. For 1910 and 1920 recorded salary workers and salaries are composites of listed variables. This means that if any of the listed fields were incorrectly transcribed my recorded values will also be wrong. The 1940 data recorded some industrial activity in a way that made it impossible to use the recorded sums to verify values for the number of salaried workers, wage earners, and establishments. In 1940, the Census categorized a wide swath of industrial activity as other and then provided information on some, but not all of the other activity. Because of this it was not possible to reconcile recorded and listed sums for these three variables. Upon reviewing many of these differences there is no evidence to believe my values or the listed values in the Census were consistently incorrect.

Each of the above cases indicate the need for researchers to be cautious when determining if these data are appropriate for their research design. Issues with the 1900 data are more serious and researchers should be cautious relying on the 1900 data as anything more than a supplement

or robustness check. Whether an issue with the original documents or the way the documents were scanned, many threes and eights look identical in the 1900 census. When reconciling the values to the recorded sums efforts were made to re-record some threes and eights to better match the listed sums. However, there many scenarios where re-recording would be little better than guessing and in this cases the original recorded values were kept.