# Constructing Historical U.S. State and County Income Inequality Data from Industrial Statistics

Jaehee Choi* James Galbraith†

University of Texas Inequality Project
Working Paper No. 79

August 2026

**Abstract**

The study of income inequality has expanded across multiple disciplines, yet long panel data that are spatially granular and temporally dense remain surprisingly scarce for the United States. This paper shows how to construct such a panel of county-level industrial earnings inequality that spans 1969 to 2022, entirely from publicly available, and largely underutilized, industrial statistics. Our approach leverages the decomposability of the Theil index, an entropy-based inequality measure: total inequality separates exactly into between-group and within-group components at any level of aggregation, without relying on individual-level microdata. The same property ensures that the county-level measures aggregate exactly into measures of earnings inequality at both the state and the national level. Combined with a concordance across industrial classification systems, the approach offers a framework for producing long, consistent estimates of income inequality with geographical depth.

*Duke Kunshan University. Email: jaehee.choi@duke.edu

†University of Texas at Austin. Email:galbraith@mail.utexas.edu

## Background & Summary

Topics in political economy often require income-inequality data that are temporally dense and spatially granular. Yet such long-panel data remain surprisingly scarce for the United States. The University of Texas Inequality Project (UTIP) has produced historical data on industrial earnings inequality going back to the 1960s for 154 countries, including the 50 U.S. states (Galbraith and Hale 2009; Galbraith et al. 2014; Galbraith 2022). This paper pushes that approach to a finer spatial scale: the U.S. county. We show that the decomposability of an information-theoretic inequality measure, combined with a concordance across industrial classification systems, makes it possible to construct a long and detailed panel of county-level industrial earnings inequality that spans 1969 to 2022, entirely from publicly available industrial statistics. The same decomposability ensures that the county-level measures aggregate into consistent measures of earnings inequality at both the state and the national level.

In theory, the ideal measure of income inequality would track the *same* income concept for *every* individual at regular intervals. In practice, this is infeasible. Some aggregation into households, tax units, or other reporting units is inevitable. Even so, data need to be collected on a consistent income concept, with good population coverage, at regular intervals. Survey data suffer from what Lustig (2020) calls the "missing rich" problem: high-income individuals are imperfectly captured because of undercoverage, sparse sampling, nonresponse, underreporting, and top-coding. In the Current Population Survey, for example, the Census Bureau's 1995 replacement of fixed top codes with cell-mean values mechanically increased measured inequality (Jenkins et al. 2011; Burkhauser et al. 2011; Burkhauser et al. 2012). These studies use restricted internal CPS data to assess the effects of top-coding.

Tax records provide greater granularity at the top of the distribution, but access is generally restricted. Changes in tax law and reporting behavior can produce artificial jumps in the resulting inequality estimates (Larrimore et al. 2021). For example, the Tax Reform Act of 1986 lowered the top individual income-tax rate below the corporate rate, which induced businesses to convert from C-corporation to pass-through S-corporation status. This shifted business income onto individual tax returns. Because active S-corporation owners also receive a substantial portion of their business income as wages, the expansion of S-corporation activity mechanically increased both reported wage income and measured top income shares. Auten and Splinter (2024) adjust for this income shifting, as well as for changes in the tax base, tax-unit composition, and income sources missing from tax returns, which lowers the top income shares measured by Piketty et al. (2018).

### Theil's (1967) "aggregation problem"

The Theil index, an information-theoretic metric, was developed by Theil (1967), who saw an immediate connection with entropy (i.e., the expected value of "surprise," or the information content of an event given the probability distribution of events), since income inequality fundamentally concerns the distribution of *income shares within a population*, which, like probabilities, are

nonnegative and sum to one (Theil 1990): How surprised should we be that a randomly drawn dollar of total income belongs to a particular individual when we already have some idea about that person's share of the population (Kanbur 2024)?

Building on this concept, Theil developed the index to compare the observed income distribution to a perfectly egalitarian one, in which income shares are exactly proportional to population shares:

$$\sum_{i=1}^{N} y_i \ln \frac{y_i}{\frac{1}{N}} = D\left(y_i \,\|\, \frac{1}{N}\right) \tag{1}$$

where $i$ denotes a unit, which can be either an individual or a group; $y_i$ represents the income share of that unit relative to total income and must therefore be nonnegative; and $\frac{1}{N}$ denotes the population share of that unit when $N$ is the total population. Theil characterized the population share as the prior and the income share as the posterior: given that perfect equality (i.e., $\frac{1}{N}$) was initially assumed, how much surprise arises when one is presented with the actual distribution (i.e., the $y_i$ values)? Hence, the Theil index represents relative entropy, formally expressed as the Kullback–Leibler divergence, $D(\cdot)$.

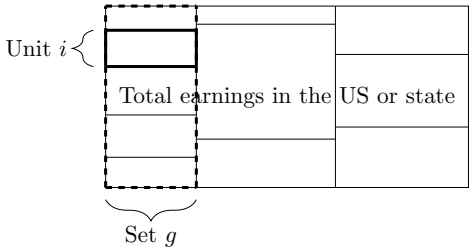



Figure 1: Total income divided into $G$ sets, with each set consisting of multiple units

Figure 1 visualizes Theil's original setup: we can think of a country's (or state's) total income as distributed among $N$ mutually exclusive units. These *units* can, in turn, be partitioned into $G$ distinct *sets*—Theil's original term for what we call *groups* below. The area of each cell represents the share of total income held by that unit.

Theil's aggregation problem arises when units are observed only as members of groups. Inequality calculated from groups' total income and population captures inequality *between* groups but omits inequality among units *within* those groups. Theil's decomposition identifies this omitted component exactly: total unit-level inequality equals between-group inequality plus an income-share-weighted sum of within-group inequality. We formalize this decomposition below.

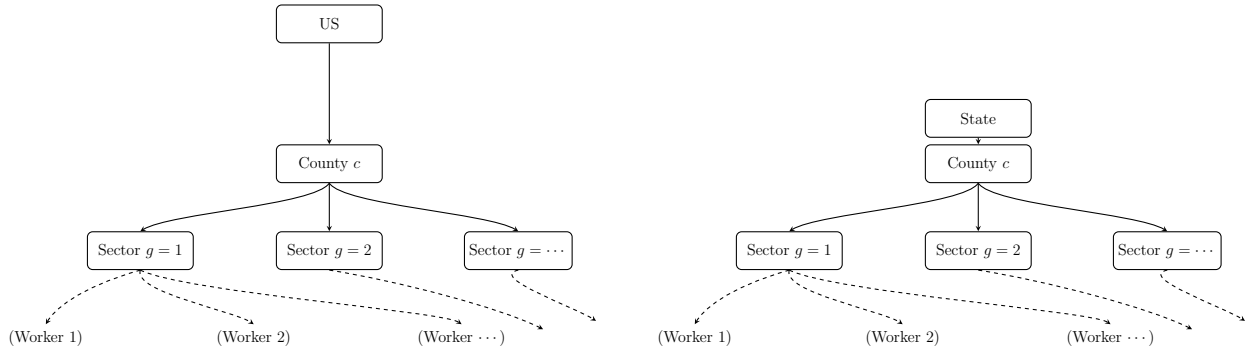



Figure 2: Aggregation occurs at two levels: first by sector, then by county

This decomposability mirrors the hierarchical structure of publicly available industrial statistics, which provide a particularly natural but underutilized application. Here, the observed units are county–sector cells in which the earnings and employment of individual workers have been summed (see Figure 2). Cell $i$ contains $n_i$ workers, and the cells differ substantially in size. Because workers are embedded within county–sector cells—the units are now divisible—we adjust the left-hand side of the original formulation in Equation 1 by replacing the uniform share $1/N$ with the cell's employment share, $q_i = n_i/N$:

$$T^* = \sum_{i=1}^{I} y_i \ln\left(\frac{y_i}{n_i/N}\right). \tag{2}$$

This adjustment preserves the egalitarian benchmark at the level of the underlying workers. If every worker earned the same amount, a cell containing $n_i$ of the economy's $N$ workers would receive $n_i/N$ of total earnings. Assigning every cell the share $1/I$ instead would imply equality across cells, regardless of their employment, rather than equality across workers. Thus, $T^*$ compares each cell's earnings share with its employment share. It is zero when average earnings are identical across all county–sector cells and increases as earnings become concentrated in cells whose employment shares are comparatively small.

Grouping the county–sector cells by county gives

$$\sum_{g=1}^{G}\sum_{i\in S_g} y_i \ln\left(\frac{y_i}{n_i/N}\right) = \underbrace{\sum_{g=1}^{G} Y_g \ln\left(\frac{Y_g}{N_g/N}\right)}_{\text{Between-county inequality}} + \underbrace{\sum_{g=1}^{G} Y_g \left\{\sum_{i\in S_g}\frac{y_i}{Y_g}\ln\left(\frac{y_i/Y_g}{n_i/N_g}\right)\right\}}_{\text{Within-county inequality}}. \tag{3}$$

where $Y_g$ is county $g$'s share of earnings, $N_g/N$ is its share of employment in the parent economy, and $S_g$ is the set of county–sector cells belonging to county $g$; cells can be summed at either the U.S. or the state level. The between-county term compares each county's earnings share with its employment share. A county's contribution is positive when its share of earnings exceeds its share of employment, negative when the reverse is true, and zero when its average earnings equal the average for the parent economy. Summing these county contributions produces between-county inequality.

The inner term in the second component measures between sectors inequality within county $g$. It compares each sector's share of county earnings with its share of county employment and is large when average earnings differ substantially across sectors within the county. Multiplying this measure by $Y_g$ gives the county's contribution to within-county inequality in the parent economy.

The decomposition therefore provides a spatial distribution of industrial earnings inequality. The between-county component identifies where earnings are concentrated relative to employment across counties, whereas the within-county component identifies where differences among industries generate earnings inequality locally. These county-level elements can then be summed to recover total inequality in the parent economy. When earnings and employment shares are calculated relative to national totals, the county contributions sum to the U.S. measure. When they are calculated relative to state totals, they sum to the corresponding state measure. As a member of the only class of inequality measures that satisfies additive decomposability (Bourguignon 1979; Shorrocks 1980), the Theil index offers great flexibility in selecting units, groups, and totals, provided the data are hierarchically nested. In Appendix A, we derive the decomposition in the same style as Theil did, contextualized to the hierarchical structure of industrial data.

Returning our earlier discussion on the ideal measure of inequality, the connection to individual inequality follows directly from Theil's aggregation problem. If individual earnings and each worker's county–sector assignment were observed, the worker-level Theil index could be written as

$$T_{\text{workers}} = T^* + \sum_{i=1}^{I} y_i T_i, \tag{4}$$

where $T_i$ measures inequality among workers within county–sector cell $i$. Our measure, $T^*$, is therefore exactly the inequality between workers across county–sector cells attributable to differences in average earnings across those cells (see Appendix A.3). The unobserved remainder is the earnings-share-weighted sum of inequality among workers employed in the same county and sector. Because this within-cell component is nonnegative, $T^*$ is a lower bound on total worker-level inequality.

The self-similar, or "fractal," property of the Theil index implies that inequality measured at different levels of aggregation displays similar movements (Conceição et al. 2001). Consistent with this property, the validation exercises presented in Section 6 show that our U.S. measures closely track established survey- and tax-based inequality series built from individual-level microdata. Our approach thus provides a long, spatially detailed, and publicly reproducible series measuring the component of earnings inequality that can be recovered from grouped industrial data.

## Methods

### Data acquisition

County-level industrial earnings and employment data are available through the Bureau of Economic Analysis (BEA)'s *Personal Income and Employment by Local Area*, under the Regional Economic Accounts. The income files are provided in two formats: `CAINC5N`, which reports personal income

by major component and earnings by NAICS industry for 2000–2022, and `CAINC5S`, which provides the same information using the SIC classification for 1969–1999. Employment data are available through two corresponding files: `CAEMP25N`, which reports total full-time and part-time employment by NAICS industry for 2000–2022, and `CAEMP25S`, which reports the same for SIC industries for 1969–1999.

Industrial statistics have largely been overlooked in the measurement of inequality. They offer an alternative to costly new surveys or often-restricted tax data, while providing broad coverage and consistency over time. In the US, state and county personal income and employment estimates are regularly compiled by various federal and state agencies. Because employer reporting to the state Unemployment Insurance system is mandatory under both federal and state law, the resulting data provide a near-complete census of employment and payroll, even at aggregated levels.

For future updates beyond 2022, the `CAEMP25N` series was discontinued in November, 2024, due to budget constraints and has been archived on the BEA website. However, it remains possible to construct county-level income inequality datasets using County Business Patterns (CBP) data by following our approach. The U.S. Census Bureau's official CBP files are available from 1986 onward on their website, and earlier files are accessible through the National Archives. Eckert et al. (2022) digitize, clean, and prepare CBP files from 1946 to 1974, but note that first-quarter payroll data are available only in sparse form prior to 1974—almost nonexistent at the county level. Raw CBP files for 1975–1985 are also available through the authors' website (Eckert et al. 2021).

When comparing raw payroll data for a single year at the SIC division level, we found that approximately 3.3% (=1,134/34,199) of county×sector payroll cells were missing in the `CAINC5S` dataset for 1976, compared to 18.5% (=5,110/27,648) in the CBP file for the same year. For 2001, at the 2-digit NAICS industry level, `CAINC5N` had 20.5% (=15,551/76,032) missing payroll data, compared to 25.4% (=16,321/60,049) in the CBP file.

Based on this comparison, we proceed under the assumption that the BEA data series requires significantly less imputation than CBP. Nonetheless, while developing consistent comparisons and crosswalks between the two datasets lies beyond the scope of this paper, it remains an important task for extending the long-run income inequality series in future research.

### Income concept

There are two distinct income concepts in the BEA data files: earnings and personal income. The income concept used in this paper is *earnings by place of work*. Earnings include wages and salaries, supplements to wages and salaries (e.g., tips, commissions, and bonuses), and proprietors' income accrued during the year. Earnings also include both employer and employee contributions to social insurance programs. It is gross of personal income taxes but net of personal contributions to government social insurance.

Earnings are measured by the place of work (i.e., the location of the employing establishment) rather than by the place of residence where the individual receives the income (Bureau of Economic Analysis 2024). Personal income, by contrast, is a broader measure based on place of residence. In

addition to earnings, Personal Income includes income from sources such as business ownership and financial assets. However, personal income is available only as an aggregate at the county level.

### Data Construction Roadmap

The construction of our dataset proceeds in three broad steps:

- Step 1: Imputation of missing employment and earnings values for all four files: `CAINC5S`, `CAINC5N`, `CAEMP25S`, and `CAEMP25N`. Hereafter, the SIC-based files are referred to as `S` and the NAICS-based files as `N`.
- Step 2: Using the crosswalk method developed by Fort and Klimek (2016), based on the 1997 Economic Census, we build a crosswalk for employment and earnings to account for the transition from the SIC to NAICS industrial classification.
- Step 3: Compute Theil components and the overall Theil index for 1969–2022 using the harmonized panel of industrial earnings and employment.

### Step 1: Imputing Missing Employment and Earnings in the BEA Files

Approximately 4.5% (= 49,186/1,093,906) of county×industry cells in the `S` files and 23.4% (= 336,029/1,437,933) in the `N` files are suppressed for confidentiality and marked "(D)". The substantially higher suppression rate in the `N` files likely reflects the doubling of two-digit sectors, which increases disclosure risk as industries become more finely segmented. Suppression is symmetric so that when earnings are withheld for a given county×industry cell, employment for that cell is also missing.

Nondisclosure typically occurs when a county×year×industry cell contains so few firms that publishing its value could reveal confidential information about a specific establishment. Although aggregate employment and earnings at the county×year level are always reported, the suppression scheme prevents unique recovery of cell-level values by simple subtraction from these totals. The nondisclosure practice therefore generates an underdetermined system of equations; for any county–year more than one industry is suppressed by design.

We impute missing sectoral employment and earnings using historical sectoral shares. This approach is motivated by the fact that, unlike the CBP data (Eckert et al. 2021), the BEA files do not provide cell-specific lower and upper bounds from suppression flags; the only hard constraints are state-level totals. We therefore assume that, within a county, the relative distribution across suppressed sectors does not change abruptly over short horizons. We test whether this assumption is reasonable in Appendix B. The approach performs especially well for the SIC regime data and less so for NAICS, but it remains the most practical way to make use of the available aggregate totals.

In practice, for each county and each contiguous block of years with an identical set of missing sectors, we identify the nearest year in which the county reports a complete set of non-missing sectoral employment or earnings values. We treat identical missing-sector configurations as a single

spell because they likely reflect the same informational constraints. This approach also improves computational efficiency.

In the nearest year, we compute each missing sector's share relative to the total employment or earnings of all sectors missing in that block. For each county–year within the block, we compute the residual to be allocated across the missing sectors by subtracting the observed sectoral values from the county total. We then distribute this residual proportionally across the missing sectors using the shares from the nearest year. For the SIC regime, this procedure imputes approximately 91% of missing values.

Imputing missing earnings presents additional complications relative to employment because earnings may be negative, whereas employment cannot. Proprietors' income can be negative when business losses exceed wages, salaries, and supplements. Negative earnings create a problem because the Theil index is defined over non-negative income shares. A common practice is to replace negative incomes with zero or a small positive value (Kim et al. 2024). We follow this convention: whenever reported earnings are non-positive, or when the sum of known sectoral values exceeds total county earnings, we replace the affected value with 1.

In summary, the imputation procedure for the SIC years (1969–2000) proceeds as follows.

1. We first identify contiguous years during which a county has the same set of missing sectors. Once these county-year blocks are defined, we search both backward and forward to locate the nearest year with a complete set of non-missing sectors. If two years are equally distant, we select the earlier one. For each missing sector in a given county–year, we calculate its share of employment and earnings relative to the total for all missing sectors in the nearest year. These totals are obtained by subtracting the sum of the missing sectors from the county aggregates. We then apply the sector-specific shares to reallocate the county–year aggregates across the missing sectors.

2. For each county, we first identify contiguous years during which the same set of sectors is missing. We treat each such county–year sequence as a single imputation block. For each block, we search both backward and forward to locate the nearest year in which all sectors are observed. If two candidate years are equally distant, we select the earlier one.

3. We interpolate with a linear trend for missing spell gaps of up to 3 years.

4. Step 1 is repeated.

5. Any values still missing are filled by evenly distributing the remaining unknown totals (i.e., averages) across the number of missing sectors for that county–year.

For the `N` regime years 2001–2022, we apply the same Steps 1–4, followed by an additional round of interpolation and averaging. We use a more relaxed interpolation window of six years rather than three because identifying a nearby year with a complete set of non-missing county×sector observations becomes more difficult once industries are doubly segmented under NAICS. Whereas

Step 1 alone accounts for approximately 91% of missing values in the `S` files, it accounts for only about 11% in the `N` files.

1. For each missing sector in a given county–year, we compute that sector's share of employment and earnings relative to the total for all missing sectors in the nearest reference year and apply those shares to the missing county–year cells.
2. We interpolate using a linear trend for missing spell gaps of six years or less.
3. Steps 1–2 are repeated.
4. Any values still missing are filled by evenly distributing the remaining unknown totals (i.e., averages) across the missing sectors for that county–year.
5. Interpolation is repeated, and averages are applied again.

### Step 2: Concordance among SIC and NAICS industry classifications

When U.S. statistical agencies transitioned from the Standard Industrial Classification (SIC) system to the North American Industry Classification System (NAICS) in the late 1990s, the change introduced a sharp discontinuity in sectoral variables such as output, value added, payroll, and employment—even when there was little change in underlying economic fundamentals. To reconcile the two systems, transition matrices have been constructed to reweight variables (Yuskavage 2007). We follow Fort and Klimek (2016), who use the 1997 Economic Census as a bridge between the two classification systems; the census reports detailed industry classifications for the same establishments under both SIC and NAICS. They apply this mapping to establishments in the Longitudinal Business Database to tabulate employment- and payroll-based weights that allow data classified under the 1987 SIC system to be converted to 1997 NAICS definitions. They also provide a concordance between 1987 SIC codes and 1997 NAICS codes.

Table 1: Concordance between NAICS Industry Classification and Concordance with SIC codes

| 2-Digit Code | Description | SIC Sectors (Division Code) |
|---|---|---|
| 111–112 | Farm employment; Farm earnings | Farm (01–02) |
| 113–115 | Forestry, fishing, and related activities | Forestry, fishing, and related activities (07–09) |
| 21 | Mining, quarrying, and oil and gas extraction | Mining (B) |
| 22 | Utilities | Transportation, communications, utilities (E) |
| 23 | Construction | Construction (C) |
| 31–33 | Manufacturing | Manufacturing (D) |
| 42 | Wholesale trade | Wholesale trade (F) |
| 44–45 | Retail trade | Retail trade (G) |
| 48–49 | Transportation and warehousing | Transportation, communications, utilities (E) |
| 51 | Information | Services (I) |
| 52 | Finance and insurance | Finance, insurance, and real estate (I) |
| 53 | Real estate and rental and leasing | Finance, insurance, and real estate (H) |
| 54 | Professional, scientific, and technical services | Services (I) |
| 55 | Management of companies and enterprises | Services (I) |
| 56 | Administrative and support and waste management and remediation services | Services (I) |
| 61 | Educational services | Services (I) |
| 62 | Health care and social assistance | Services (I) |
| 71 | Arts, entertainment, and recreation | Services (I) |
| 72 | Accommodation and food services | Services (I) |
| 81 | Other services (except government and government enterprises) | Services (I) |
|  | Government and government enterprises | Government and government enterprises |

**Note:** This concordance maps 2-digit NAICS industry codes to SIC divisions based on Economic Analysis (2024). There are a total of eleven SIC divisions and 21 2-digit sectors under NAICS. This is the most aggregated level available in both the `CAINC5N/S` and `CAEMP25N/S` files; while earnings are available at a more disaggregated 3-digit level, employment is available only at the 2-digit/division level.

The 11 broad SIC-based sector groups shown in Table 1 define the common two-digit level of industry aggregation. We therefore consolidate the 21 NAICS-based industry categories into these 11 broader groups to make concordance across the two classification systems more tractable. The reference classification used to harmonize the series over time, however, is the 1997 NAICS system. We convert earlier SIC-era earnings and employment into the corresponding 1997 NAICS definitions before combining them with the later NAICS-era data. Thus, the harmonized series is organized into 11 broad, two-digit sector groups, while its values are expressed on a common 1997 NAICS basis.

In practice, we first use the BEA industry categories shown in Table 1 to organize the SIC-era observations into the 11 broad, two-digit sector groups. Second, we use the transition weights publicly provided by Fort and Klimek (2016) to convert these SIC-based values to a 1997 NAICS basis. At the two-digit level, the concordance can be applied directly: each observed SIC-based sector is allocated across one or more NAICS-based sectors using the published transition weights, without requiring us to infer unobserved child industries from their parent-industry totals. Extending the concordance to the three-digit level would require such imputation. We therefore retain the two-digit level. Finally, separate weights are used for employment and earnings: employment-based weights are applied to employment, while payroll-based weights are applied to earnings.

As an example of the employment concordance, approximately 99.7% of employment classified under SIC construction is assigned to NAICS construction, while the remaining 0.3% is assigned to NAICS services. NAICS construction also receives 0.57% of employment classified under SIC finance, insurance, and real estate and 0.40% of employment classified under SIC services. Because

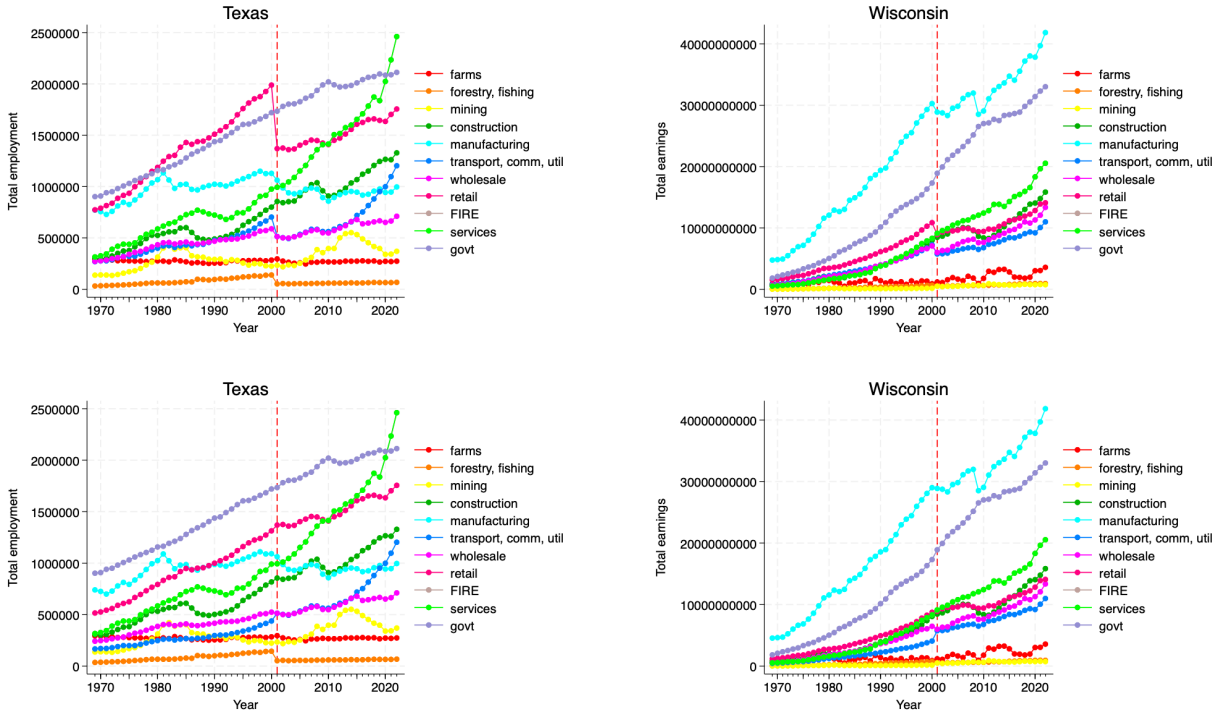
these rates are calculated relative to different SIC source sectors, the rates contributing to NAICS construction need not sum to 100%. Instead, for each SIC source sector, the allocations across all NAICS destination sectors sum to 100%. The employment-based weights therefore preserve total employment. The payroll-based weights satisfy the same condition, thereby preserving total earnings.

Figure 3 illustrates the effect of SIC-to-NAICS reweighting. The left panel (Figure 3a) shows employment by sector in Texas after reweighting, while the right panel (Figure 3b) shows earnings by sector in Wisconsin after reweighting. In both panels, the sharp break associated with the classification change is visibly reduced. A limitation of this approach is that transition weights derived from the 1997 national establishment data are applied uniformly across counties and years. The procedure therefore assumes that the allocation of each SIC sector across NAICS destination sectors does not vary geographically or over time.

Figure 3: Effect of reweighting SIC to NAICS in employment and earnings series

(a) Texas employment—before reweighting (top) and after reweighting (bottom)

(b) Wisconsin earnings—before reweighting (top) and after reweighting (bottom)




**Note:** Each panel shows the series before (top) and after (bottom) reweighting using the method developed by Fort and Klimek (2016). The red vertical line indicates the transition from SIC to NAICS classification.

## Step 3: Computation of the Theil index

Before constructing the indices, we exclude county×sector cells with positive earnings but zero employment. In such a cell, the earnings share is positive, $y_i > 0$, while the employment share is zero, $q_i = 0$. Its contribution, $y_i \ln(y_i/q_i)$, is therefore undefined and diverges to $+\infty$ as $q_i$ approaches zero from above. These cells generally contain only small positive earnings amounts and account for 0.57% of the original sample (= 10,577/1,847,109).

After imposing this support condition, we compute the county-level terms in Equation 1 and aggregate them to obtain the between-county component, the within-county component, and the total Theil index for each parent economy. Table 2 lists the constructed variables, their corresponding mathematical terms in Equation 1, and their interpretations.

Three variables retain county-level information: `between`, `theil_within1`, and `within2`. Figures 4, 6, and 5 map these variables at roughly decennial intervals from 1969 through 2019, with an additional map for 2022, the final year available in this version of the data. The variable `between` is each county's signed contribution to the between-county component. The maps use a diverging blue-to-red scale. Darker red denotes larger positive contributions, with the darkest red identifying the upper 0.1% of the distribution. The remaining warm colors identify other relatively large positive

contributions within the upper 10%. Shades of blue denote negative contributions, with darker blue indicating more negative values. These negative terms offset positive county contributions in the aggregate, although the aggregate between-county component remains nonnegative; see Appendix A.5.

The upper tail of the positive between-county contributions is well approximated by a power law with a Zipf-like exponent. Across the seven map years, Hill and log-rank estimates of the Pareto exponent range from 0.9 to 1.5 and are centered near 1.1, the same general exponent class observed in city- and firm-size distributions (Gabaix 2009). The estimated power-law regime spans roughly the upper 1% of counties. An exponent this close to one indicates that the positive contribution mass is concentrated among a handful of counties rather than spread broadly across the upper tail.

The county rankings from Table 3 reflect this concentration. Within each map year, we rank counties by their between-county contribution from largest to smallest; among the roughly 3,110 counties, the sixteenth-largest value marks approximately the boundary of the upper 0.5%. New York County, New York is the largest contributor in all seven map years, with a contribution between 11.5 and 17.5 times this sixteenth-largest value. Summing down the ranking, the four largest counties account for one-third to two-fifths of the sum of all positive between-county contributions, and the sixteen counties of the upper 0.5% for roughly two-thirds, a share that reaches 73% in 2022. Membership in the extreme upper tail is also persistent: only nine distinct counties occupy a top-four position across the seven map years. Its composition nevertheless shifts from the large industrial metropolitan counties of 1969–1979, including New York, Los Angeles, Cook, and Wayne, toward counties and county-equivalent units associated with technology, finance, government, and energy, including Santa Clara, San Francisco, King, the District of Columbia, and Harris, by the end of the sample.

The variable `theil_within1` measures between-sector earnings inequality within a county before the county is weighted by its share of earnings in the parent economy. The variable `within2` multiplies this county-specific index by the county's earnings share and therefore measures the county's contribution to the aggregate within-county component. Both variables are nonnegative, so their maps use a sequential yellow-to-red scale, with darker shades indicating greater inequality.

The upper tails of the two within-county measures differ sharply, and the difference comes almost entirely from the earnings-share weight. Because county earnings shares vary over several orders of magnitude, `within2` does too: across map years the largest value of `within2` is 200 to 1,070 times the median, and Pareto exponents of roughly 1.4 to 2.6 indicate a heavy tail, with the largest economies—Harris, New York, Los Angeles, and Denver—at the top (see Table 4). The tail is nevertheless less extreme than that of `between`, because the two factors of $\texttt{within2}_g = Y_g \times \texttt{theil\_within1}_g$ are negatively correlated across counties: the largest counties are sectorally diversified, with below-median values of `theil_within1` (e.g., New York's 2022 value is 0.035, against a cross-county median of 0.074).

The variable `theil_within1` itself carries no earnings-share weight, and its tail is much more compressed: for county $g$ its maximum is $\ln(N_g/n_{g,\min})$, attained when all of the county's earnings

accrue to the sector with the smallest employment $n_{g,\min}$ (see Appendix A.6.2). In Figure 6, the maximum is only 11 to 18 times the cross-county median, extreme values are only 1.3 to 1.7 times the top-0.5% cutoff, and exponent estimates of 3 to 5.5 describe a thin tail. The membership of the upper tail is also unstable: the 28 top-four positions across the seven map years are held by 23 different counties, and only four counties ever appear in more than one year (see Table 5). The top counties are small rural counties whose earnings happen to be dominated by one or two sectors in a given year in contrast to the persistent membership of the `between` tail. These differences motivate the binning of the county maps: the maps of `between` add bins in the extreme upper tail, including a separate top-0.1% category whose members enter at five to ten times the value marking the top 1%, whereas the within-county maps terminate at the top 0.5%.

At the parent-economy level, summing `between` across counties produces the between-county component, while summing `within2` produces the within-county component. At the national level, the resulting variables are `theil_between`, `theil_within`, and `theil_total`. Their state-level counterparts, constructed using each state as the parent economy, are `theil_between_s`, `theil_within_s`, and `theil_total_s`. Figure 7 presents these series for the United States, New York, and California. The shaded areas represent the between- and within-county components, while the total series reports the corresponding total Theil index. The complete set of state-level time series is reported in Appendix C. Because the absolute magnitude of the Theil index does not have an immediately intuitive interpretation, we interpret these measures primarily through comparisons across places and over time.

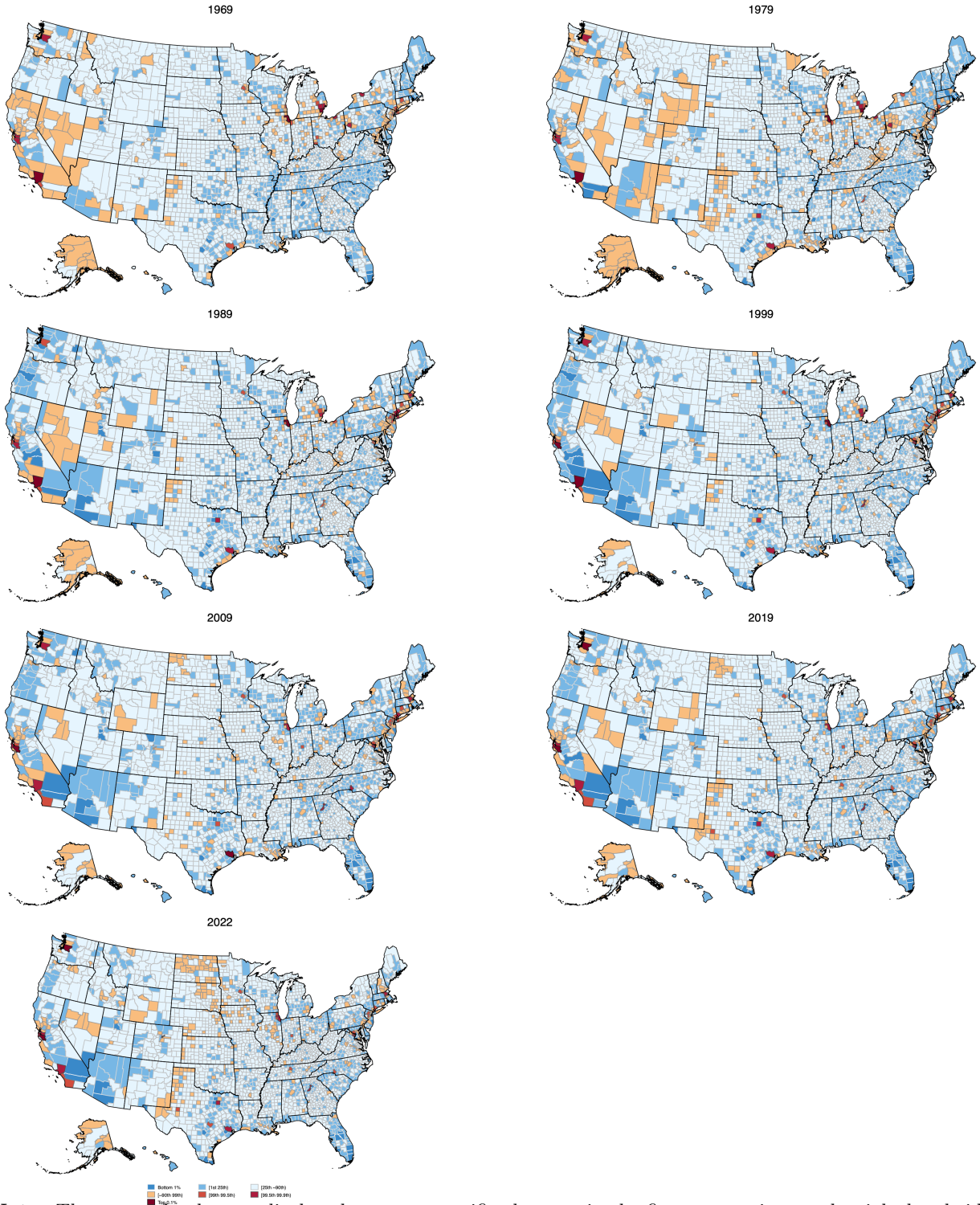
For the purposes of this paper, we do not provide an extensive analysis of either the spatial distribution of the county-level contributions to between- and within-county inequality or the evolution of the corresponding aggregate measures. Both are left for future work.

Table 2: Variable codebook for the `theil_harmonized` data

| Variable name | | | |
|---|---|---|---|
| **U.S. parent** | **State parent** | **Formula** | **Description** |
| `theil_total` | `theil_total_s` | $T^* = \sum_{i=1}^{N} y_i \ln \frac{y_i}{n_i/N}$ | Total Theil index at the U.S. or state parent level, equal to aggregate between-county inequality plus aggregate within-county inequality. |
| *Additive decomposition of $T^*$* | | | |
| `between` | `between_s` | $Y_g \ln \frac{Y_g}{N_g/N}$ | County $g$'s contribution to parent-level between-county inequality. |
| `theil_between` | `theil_between_s` | $\sum_{g=1}^{G} Y_g \ln \frac{Y_g}{N_g/N}$ | Parent-level between-county inequality, summed over counties and constant across counties within each parent-year. |
| `within1` | Same as `within1` because $\frac{y_i}{Y_g}$ and $\frac{n_i}{N_g}$ are invariant to the choice of parent economy | $\frac{y_i}{Y_g} \ln \frac{y_i/Y_g}{n_i/N_g}$ | Sector $i$'s contribution to between-sector inequality within county $g$, before applying the county earnings-share weight $Y_g$. |
| `theil_within1` | Same as `theil_within1` | $\sum_{i \in S_g} \frac{y_i}{Y_g} \ln \frac{y_i/Y_g}{n_i/N_g}$ | County $g$'s between-sector inequality, summed over sectors but not yet weighted by $Y_g$; the county-specific within-county inequality before applying the county earnings-share weight. |
| `within2` | `within2_s` | $Y_g \sum_{i \in S_g} \frac{y_i}{Y_g} \ln \frac{y_i/Y_g}{n_i/N_g}$ | County $g$'s earnings-share-weighted contribution to parent-level within-county inequality, equal to $Y_g \times$ `theil_within1`. |
| `theil_within` | `theil_within_s` | $\sum_{g=1}^{G} Y_g \sum_{i \in S_g} \frac{y_i}{Y_g} \ln \frac{y_i/Y_g}{n_i/N_g}$ | Parent-level within-county inequality, summed over counties and constant across counties within each parent-year. |

*Note:* Unit $i$ indexes an industrial sector and $g$ indexes a county; $S_g$ is the set of sectors in county $g$, $G$ the number of counties, and $N$ the total employment in the parent population. $y_i$: the earnings income share of sector $i$ relative to total U.S. or state earnings income, depending on the parent level chosen. $Y_g$: the earnings income share of county $g$ relative to total U.S. or state earnings income, depending on the parent level chosen. $n_i$ and $N_g$ denote employment in sector $i$ and in county $g$. Columns under "U.S. parent" normalize every share to the entire U.S.; the "State parent" variables (suffix `_s`) are defined identically but normalize to the state. The summed components (`theil_between`, `theil_within`) and the total (`theil_total`) are identical across all counties within a year; their state-parent counterparts (`theil_between_s`, `theil_within_s`, `theil_total_s`) are identical across counties within the same state.

Figure 4: Evolution of county-level inequality, 1969–2022: between-county contributions




**Note:** The county-level maps display the county-specific elements in the first summation on the right-hand side of Equation 1, $Y_g \ln\left(\frac{Y_g}{N_g/N}\right)$ (denoted by `between`), where $Y_g$ is county $g$'s share of total U.S. earnings and $N_g/N$ is its share of total U.S. employment. The between-county element captures each county's contribution to overall inequality: positive values (orange to red) indicate that the county increases overall U.S. inequality in that year, while negative values (blue hues) indicate that it reduces it. Within each year, counties are ranked by percentile. The orange–red spectrum represents counties that typically fall within the top 10% of the distribution: roughly the 90th–99th percentiles, the 99th–99.5th percentile, the 99.5th–99.9th percentile, and the top 0.1%. The three blue shades represent the lower portion of the distribution: the bottom 1%, the 1st–25th percentiles, and roughly the 25th–90th percentiles.

Table 3: Top 10 counties by between-county Theil contributions

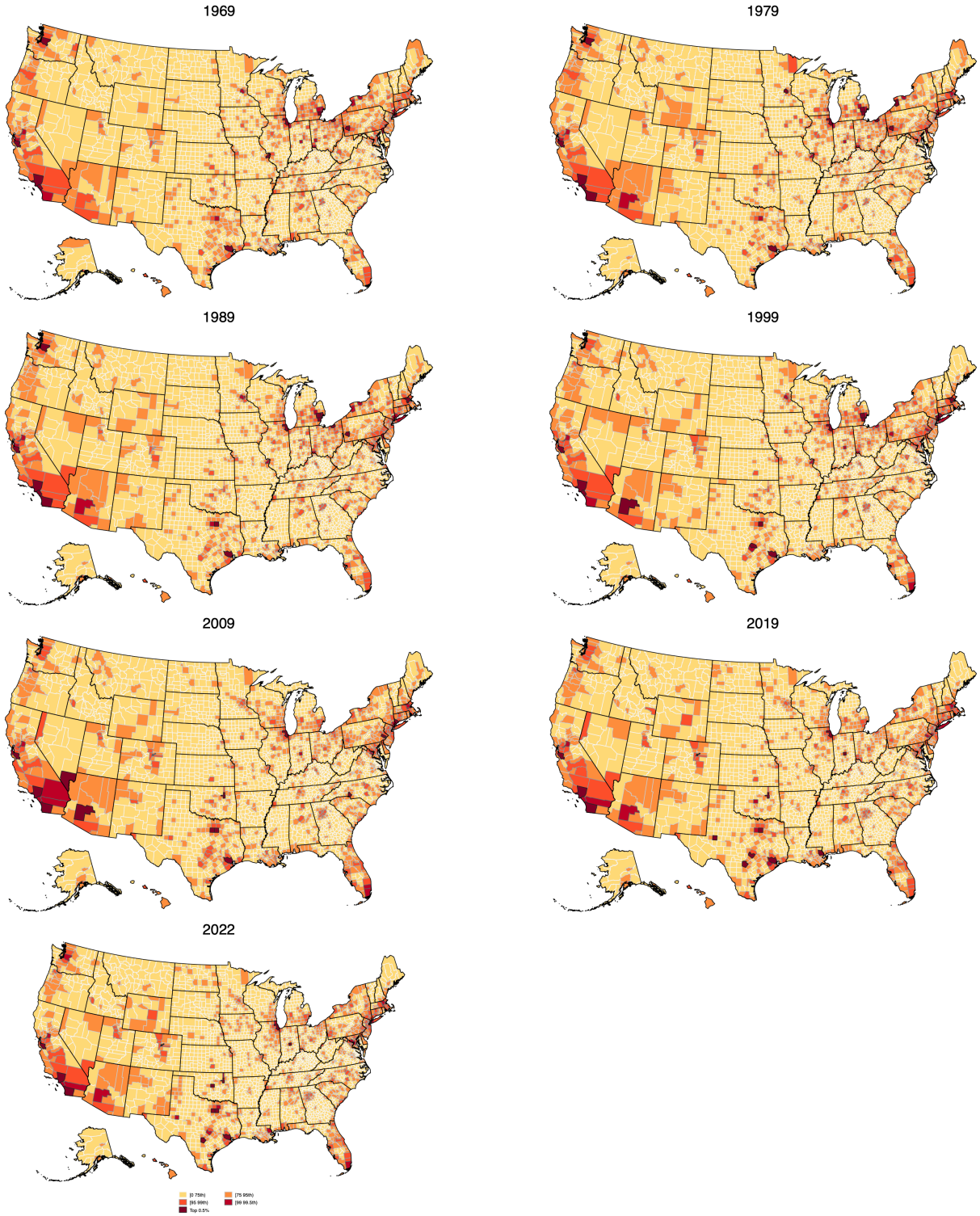
| | 1969 | 1979 | 1989 |
|---|---|---|---|
| 1 | New York, NY | New York, NY | New York, NY |
| 2 | Los Angeles, CA | Cook, IL | Los Angeles, CA |
| 3 | Cook, IL | Los Angeles, CA | Cook, IL |
| 4 | Wayne, MI | Wayne, MI | District of Columbia, DC |
| 5 | District of Columbia, DC | District of Columbia, DC | Santa Clara, CA |
| 6 | San Francisco, CA | Harris, TX | Wayne, MI |
| 7 | Cuyahoga, OH | San Francisco, CA | San Francisco, CA |
| 8 | Philadelphia, PA | Cuyahoga, OH | Suffolk, MA |
| 9 | Oakland, MI | King, WA | Harris, TX |
| 10 | King, WA | Philadelphia, PA | Middlesex, MA |

| | 1999 | 2009 | 2019 |
|---|---|---|---|
| 1 | New York, NY | New York, NY | New York, NY |
| 2 | Santa Clara, CA | Santa Clara, CA | Santa Clara, CA |
| 3 | Cook, IL | District of Columbia, DC | San Francisco, CA |
| 4 | Los Angeles, CA | Harris, TX | King, WA |
| 5 | Harris, TX | Fairfield, CT | District of Columbia, DC |
| 6 | King, WA | Los Angeles, CA | Suffolk, MA |
| 7 | District of Columbia, DC | Suffolk, MA | Harris, TX |
| 8 | Dallas, TX | San Francisco, CA | Los Angeles, CA |
| 9 | San Francisco, CA | Cook, IL | San Mateo, CA |
| 10 | Suffolk, MA | Philadelphia, PA | Cook, IL |

| | 2022 |
|---|---|
| 1 | New York, NY |
| 2 | Santa Clara, CA |
| 3 | San Francisco, CA |
| 4 | King, WA |
| 5 | San Mateo, CA |
| 6 | Suffolk, MA |
| 7 | District of Columbia, DC |
| 8 | Harris, TX |
| 9 | Dallas, TX |
| 10 | Middlesex, MA |

**Note:** For each year, counties are ranked from largest to smallest by their county-specific between-county Theil element, $Y_g \ln\left(\frac{Y_g}{N_g/N}\right)$ (denoted by `between`), where $Y_g$ is county $g$'s share of total U.S. earnings and $N_g/N$ is its share of total U.S. employment. The table reports the ten largest elements. Hill estimates based on the largest $k$ observations, for $k = 5, \ldots, 10$, yield Pareto exponents mostly between 1 and 1.8, with many estimates between 1 and 1.5. Together with year-specific log–log survival plots, these estimates indicate that only the extreme upper tail exhibits a very heavy Pareto-type decay. The visually identified tail contains 5–10 counties across the sample, or roughly 0.2–0.3% of all counties, and generally 5–7 counties, or about 0.2%, from 1989 onward. New York ranks first throughout and, in recent years, its between element is roughly twice Santa Clara's, indicating a clear hierarchy even within the extreme tail. Future research could directly estimate the threshold separating the main body of the distribution from the extreme tail.

Figure 5: Evolution of county-level inequality, 1969–2022: within-county contributions




**Note:** The county-level maps display each county's earnings-share-weighted within-county element, $Y_g \left\{ \sum_{i \in S_g} \frac{y_i}{Y_g} \ln\left( \frac{y_i/Y_g}{n_i/N_g} \right) \right\}$ (denoted by `within2`), where $Y_g$ is county $g$'s share of total U.S. earnings and serves as the weight on the term in braces, while $y_i/Y_g$ and $n_i/N_g$ are sector $i$'s shares of county $g$'s earnings and employment, respectively. By construction, within-county elements are non-negative, and the color spectrum runs from yellow to deep red, without a diverging scale as in the between-county series. Within each year, counties are ranked by percentile: bottom 75th, 75th–95th, 95th–99th, 99th–99.5th, and the top 0.5%.

Table 4: Top 10 counties by earnings-share-weighted within-county contributions

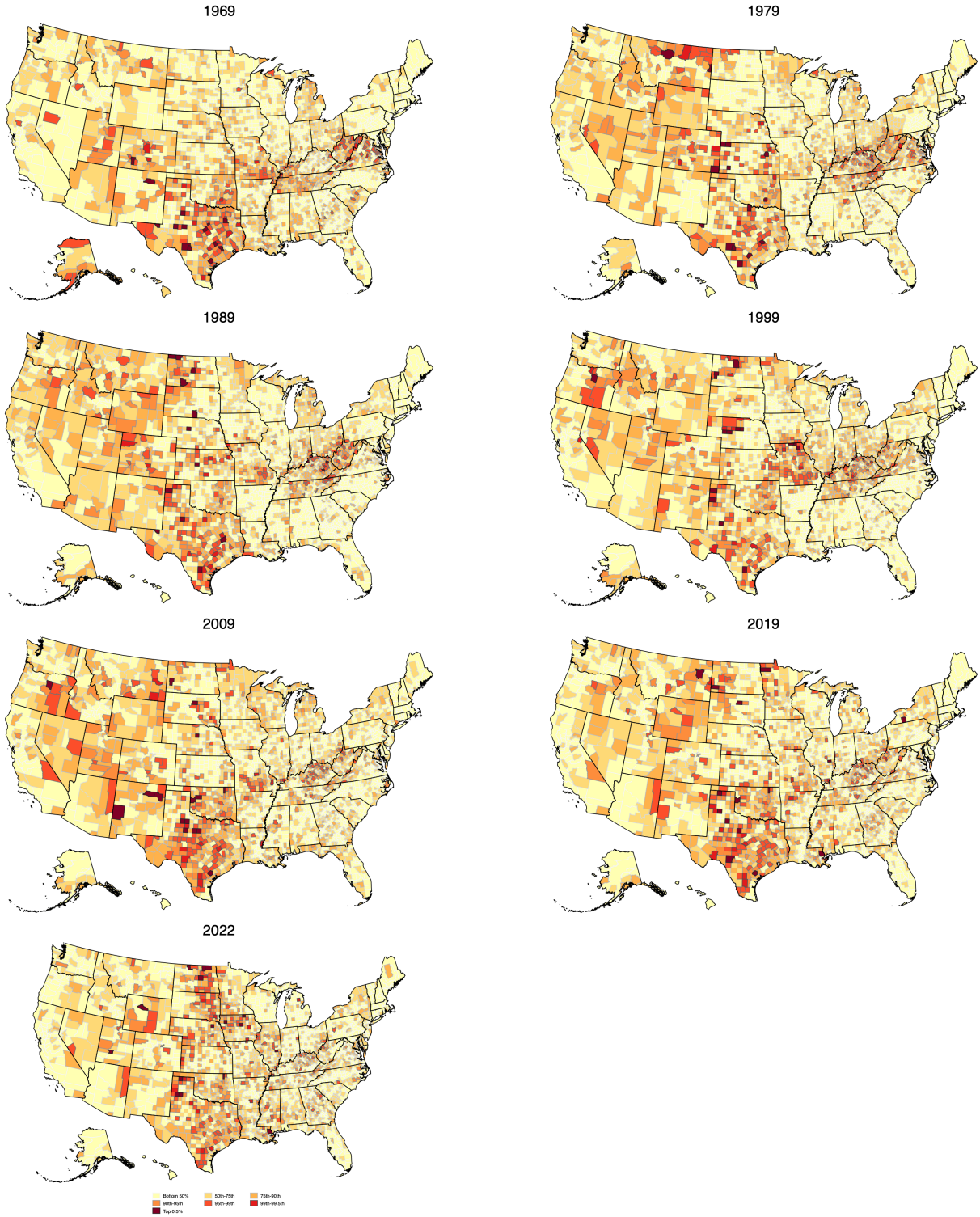
| | 1969 | 1979 | 1989 |
|---|---|---|---|
| 1 | Los Angeles, CA | Los Angeles, CA | Wayne, MI |
| 2 | Wayne, MI | Wayne, MI | Los Angeles, CA |
| 3 | Cook, IL | Cook, IL | Harris, TX |
| 4 | Cuyahoga, OH | Harris, TX | New York, NY |
| 5 | New York, NY | Allegheny, PA | Santa Clara, CA |
| 6 | Harris, TX | Cuyahoga, OH | Cook, IL |
| 7 | Allegheny, PA | New York, NY | Cuyahoga, OH |
| 8 | District of Columbia, DC | San Diego, CA | Orange, CA |
| 9 | Philadelphia, PA | King, WA | San Diego, CA |
| 10 | King, WA | Oakland, MI | Dallas, TX |

| | 1999 | 2009 | 2019 |
|---|---|---|---|
| 1 | New York, NY | Harris, TX | Harris, TX |
| 2 | Santa Clara, CA | New York, NY | New York, NY |
| 3 | Harris, TX | Fairfield, CT | Dallas, TX |
| 4 | Cook, IL | Mecklenburg, NC | Denver, CO |
| 5 | Wayne, MI | Los Angeles, CA | Los Angeles, CA |
| 6 | Los Angeles, CA | Montgomery, PA | Santa Clara, CA |
| 7 | Dallas, TX | Santa Clara, CA | Travis, TX |
| 8 | Williamson, TX | Suffolk, MA | Marion, IN |
| 9 | Suffolk, MA | Cook, IL | St. Tammany, LA |
| 10 | Fairfield, CT | San Diego, CA | Tulsa, OK |

| | 2022 |
|---|---|
| 1 | Harris, TX |
| 2 | Denver, CO |
| 3 | New York, NY |
| 4 | Santa Clara, CA |
| 5 | Dallas, TX |
| 6 | Los Angeles, CA |
| 7 | Tarrant, TX |
| 8 | Oklahoma, OK |
| 9 | Tulsa, OK |
| 10 | Marion, IN |

**Note:** The table lists, for each selected year, the top 10 counties by their earnings-share-weighted within-county elements, $Y_g \left\{ \sum_{i \in S_g} \frac{y_i}{Y_g} \ln\left( \frac{y_i/Y_g}{n_i/N_g} \right) \right\}$ (denoted by `within2`), where $Y_g$ is county $g$'s share of total U.S. earnings and serves as the weight on the term in braces, while $y_i/Y_g$ and $n_i/N_g$ are sector $i$'s shares of county $g$'s earnings and employment, respectively. Hill estimates of the Pareto exponent $\alpha$, based on the largest $k = 10, \ldots, 20$ observations, range from approximately 1.3 to 2.8 across the sample years, with most estimates between about 1.5 and 2.5. Together with year-specific log–log survival plots, these estimates suggest that only the extreme upper tail (i.e., approximately the top 0.5% of counties, or about 16 counties per year) is consistent with Pareto-type decay. The estimated exponent does not change monotonically over time. Because a lower Pareto exponent means that the tail declines more slowly, the relatively low and stable 2022 estimates, approximately 1.35–1.58, indicate that exceptionally large elements were more prominent relative to the rest of the tail in 2022 than in most earlier years. Future research could directly estimate the threshold separating the main body of the distribution from the extreme upper tail.

Figure 6: Evolution of county-level inequality, 1969–2022: within-county inequality before county earnings-share weighting




**Note:** The county-level maps display each county's unweighted within-county term, $\sum_{i \in S_g} \frac{y_i}{Y_g} \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right)$ (denoted by `theil_within1`), where $y_i/Y_g$ and $n_i/N_g$ are sector $i$'s shares of county $g$'s earnings and employment, respectively. Unlike `within2`, `theil_within1` is not multiplied by $Y_g$, the county's share of total U.S. earnings; counties are therefore represented independently of their shares of national earnings. By construction, the within-county terms are non-negative, and the color spectrum runs from yellow to deep red. Within each year, counties are ranked by percentile: the bottom 75%, the 75th–95th percentiles, the 95th–99th percentiles, the 99th–99.5th percentiles, and the top 0.5%.

Table 5: Top 10 counties by within-county inequality before county earnings-share weighting

| | 1969 | 1979 | 1989 |
|---|---|---|---|
| 1 | Loving, TX | Chouteau, MT | Loving, TX |
| 2 | Goliad, TX | Haskell, KS | Kent, TX |
| 3 | Real, TX | Loving, TX | Burke, ND |
| 4 | Carlisle, KY | Wilson, TX | Divide, ND |
| 5 | Somervell, TX | Cheyenne, KS | Elliott, KY |
| 6 | Hinsdale, CO | Real, TX | McMullen, TX |
| 7 | Glasscock, TX | Clay, TX | Owsley, KY |
| 8 | Lee, TX | Coal, OK | Grant, ND |
| 9 | Austin, TX | Wibaux, MT | Leon, TX |
| 10 | Coke, TX | Wabaunsee, KS | Rush, KS |

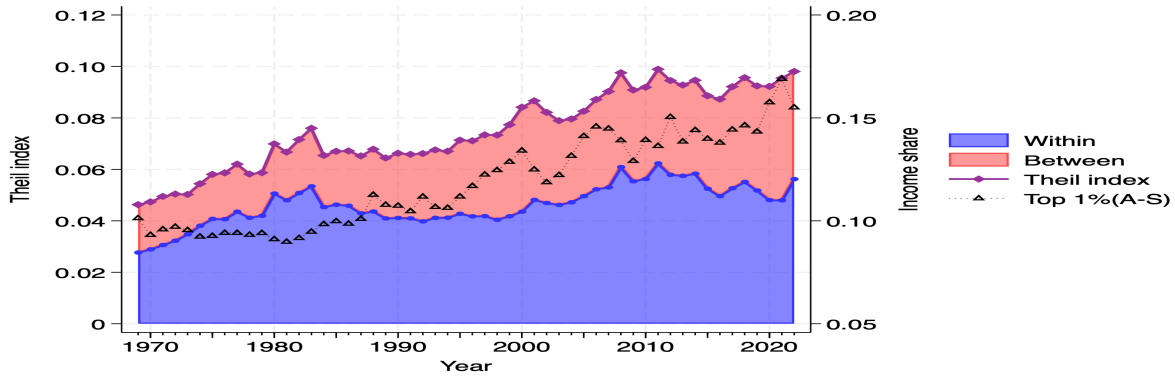
| | 1999 | 2009 | 2019 |
|---|---|---|---|
| 1 | McPherson, NE | Washington, OK | Slope, ND |
| 2 | Loup, NE | Loup, NE | Cumberland, IL |
| 3 | Borden, TX | King, TX | St. Tammany, LA |
| 4 | Arthur, NE | Wheeler, OR | Marshall, MN |
| 5 | McMullen, TX | Billings, ND | Allegany, NY |
| 6 | Blaine, NE | Goliad, TX | Sherman, TX |
| 7 | Wheeler, OR | Shackelford, TX | Washington, OK |
| 8 | McHenry, ND | Dickens, TX | McCone, MT |
| 9 | Hartley, TX | Catron, NM | Potter, SD |
| 10 | Trimble, KY | Potter, SD | Swisher, TX |

| | 2022 |
|---|---|
| 1 | Pierce, NE |
| 2 | Washington, OK |
| 3 | Cumberland, IL |
| 4 | Cherokee, IA |
| 5 | Swisher, TX |
| 6 | Hot Springs, WY |
| 7 | Renville, ND |
| 8 | Sherman, TX |
| 9 | Piute, UT |
| 10 | Mitchell, IA |

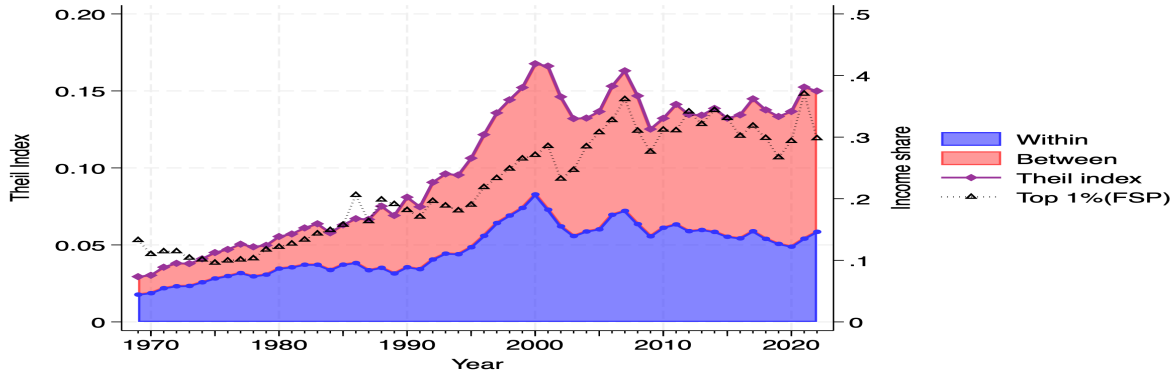
**Note:** The table lists, for each selected year, the top 10 counties by their county-specific within-county terms, unweighted by $Y_g$, $\sum_{i \in S_g} \frac{y_i}{Y_g} \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right)$ (denoted by `theil_within1`), where $Y_g$ is county $g$'s share of total U.S. earnings, while $y_i/Y_g$ and $n_i/N_g$ are sector $i$'s shares of county $g$'s earnings and employment, respectively. At the top-0.5% cutoff (i.e., approximately the 16 largest observations per year) Hill estimates of the Pareto exponent $\alpha$ range from about 3.0 to 5.8 across the sample years. These exponents are substantially higher than the corresponding `within2` estimates, which mostly range from about 1.5 to 2.5. Because a higher $\alpha$ implies faster Pareto decay, the comparison indicates that the upper tail of `theil_within1` is thinner and that extreme counties are less dominant relative to the rest of its distribution.

Figure 7: Evolution of income inequality, 1969–2022, decomposed into within- and between-county inequality

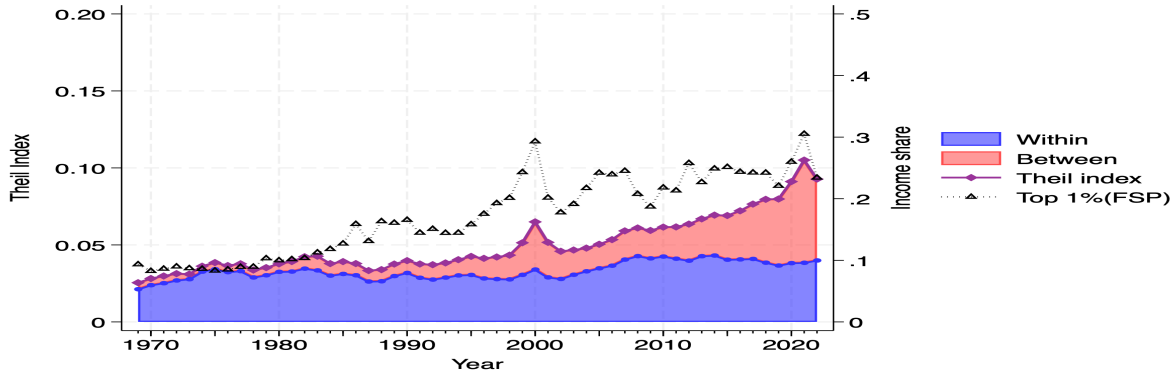
(a) US




(b) New York




(c) California




**Note:** The red diamond series reports the total Theil index, $T^*$, defined in Equation (1) as the sum of the between- and within-county components, which are shown by the red and blue shaded areas, respectively. In Figure 7a, employment and earnings shares are calculated relative to annual national totals (`theil_total`); in Figures 7b and 7c, they are calculated relative to annual state totals (`theil_total_s`). The black series plots the top 1% income share for comparison: Figure 7a uses the series from Auten and Splinter (2024), whereas Figures 7b and 7c use the Frank–Sommeiller–Price series (Frank et al. 2015). For the national comparison, we use the pretax series to align its income concept more closely with our earnings measure. For NY and CA, the Pearson correlations are 0.93 and 0.84 in levels, and 0.59 and 0.71 in annual changes, respectively; see Table A1.

## Technical Validation

For quality control, we first impose two checks on the construction of the harmonized data. First, our imputation procedure constrains imputed industry-level employment and earnings so that they do not exceed the corresponding published totals. Second, for each state, we plot total employment and earnings by industry over time to verify that the mapping from SIC to NAICS does not introduce a discontinuity around the year 2000. We then conduct two additional validation exercises.

### Internal consistency

As shown in Appendix A.3, Equation 1 implies that a correctly implemented Theil decomposition must satisfy an exact identity: the total Theil index computed directly from the left-hand side of the equation must equal the sum of its between- and within-county components. In our implementation, these decomposed totals are stored as `theil_total` at the national level and `theil_total_s` at the state level. We verify this identity for the United States as a whole and for each of the 50 states and the District of Columbia.

For every national- and state-year observation, we compare the total Theil index calculated directly from the county–sector cells with the sum of the separately calculated between- and within-county components. Specifically, we calculate the scale-adjusted absolute difference $\left|T^{\text{direct}} - (T^B + T^W)\right| / \left(\left|T^B + T^W\right| + 1\right)$, where $p$ denotes the parent economy and $t$ denotes year. This difference is less than $10^{-13}$ for every national- and state-year observation. Thus, the total Theil index calculated directly from the county–sector cells equals the sum of the between- and within-county components up to floating-point precision.

### Comparison with existing national inequality series

Table 6 compares our series over 1969–2022 with the three Auten and Splinter (2024) top 1% concepts, the updated Piketty and Saez (2003) fiscal-income series including realized capital gains (as reproduced in the AS data files), the Census equivalence-adjusted CPS Gini and Theil, and four Gini series that Burkhauser et al. (2012) construct from internal March CPS microdata: a pretax, post-transfer household-income measure and a pretax, pretransfer tax-unit (market-income) measure, each computed for the full distribution and for the bottom 99 percent, available through 2006. In levels our index is highly correlated with every benchmark: 0.89 with the AS pretax share, 0.91 with the PS fiscal-income series and the CPS Gini, and 0.87–0.89 with the four BFJL Ginis. On the long-run path of U.S. inequality, all of these series agree.

Annual first differences are the harder test, and Panel D serves as a point of reference. The Census and BFJL post-transfer household measures are computed from the same March CPS microdata under the same money-income concept and differ only in equivalence scale; their annual changes correlate 0.92. That is effectively the ceiling. Changing only the income concept and sharing unit within the same survey—the BFJL pretransfer tax-unit measure against the CPS Gini—lowers the correlation to 0.66, in part because countercyclical transfers decouple pretransfer

from post-transfer inequality year to year: in a recession, market income at the bottom collapses while unemployment insurance and other transfers expand to replace part of it, so pretransfer inequality jumps while post-transfer inequality is more stable.

Changing the data source also lowers the correlation nearly to zero: annual changes in the AS pretax top share correlate 0.05 with the CPS Theil and 0.09 with the CPS Gini. Our index differs from every benchmark in source, income concept, and sharing unit, yet its first-difference correlations of 0.10–0.22 sit at or above what cross-source comparisons among other published series deliver. Spearman rank correlations, less sensitive to large capital-gains realization years, are uniformly higher: 0.17 rather than 0.10 against the AS pretax share, 0.26 rather than 0.16 against the PS series. Realizations bunch in particular years—1986 ahead of the Tax Reform Act rate increase, for example, and 2020–2022 during the pandemic.

The BFJL bottom-99 series test the mechanism directly. If our index captures dispersion below the top percentile, it should track inequality among the bottom 99 percent—which a top-share measure cannot. It does: our Theil matches the level path of the bottom-99 Ginis (0.88–0.89) and its annual changes co-move positively with them (0.05 and 0.17), while annual changes in the top 1% share co-move negatively with bottom-99 inequality ($-0.33$ against the pretransfer tax-unit series in Panel D, $-0.48$ in rank terms).

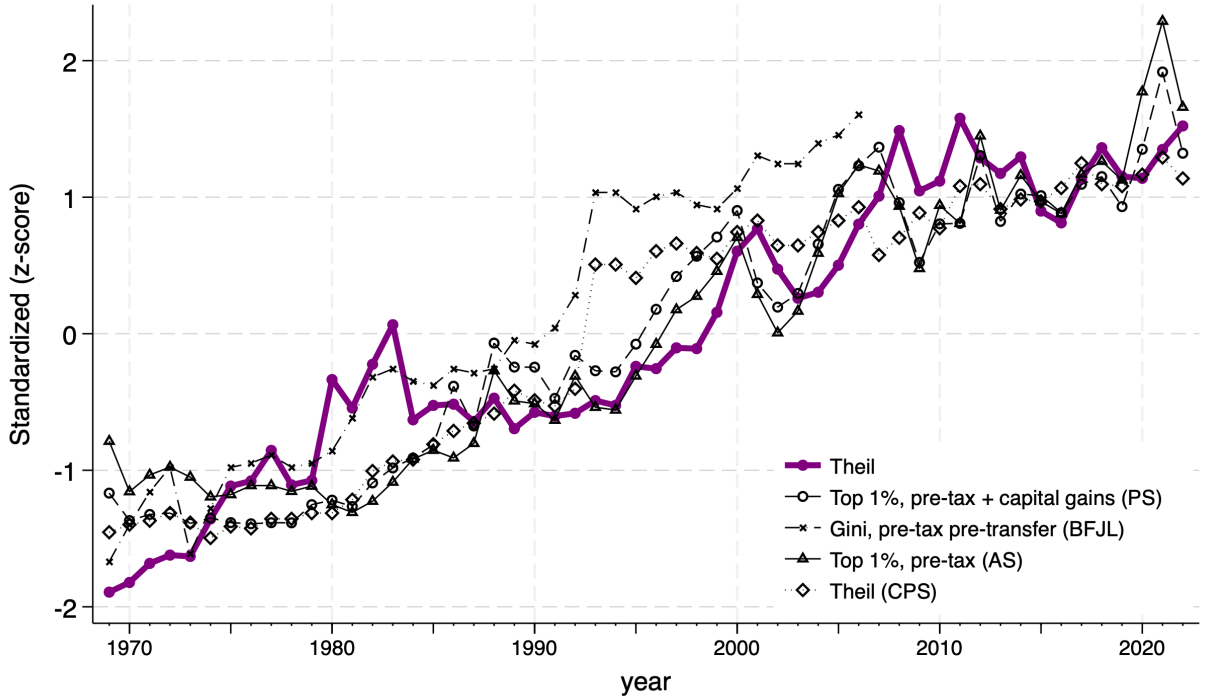
Figure 8 plots the standardized series. The common path is unmistakable; the divergences are dated and have names. In 1980–1983 our index runs above the benchmarks: the manufacturing recessions raised compensation per worker while hiring collapsed, dispersion that place-of-work earnings record and resident fiscal income does not. An exact sector decomposition of the index bears this out: manufacturing's contribution matches or exceeds the index's entire net rise above its 1975–78 level in 1980–1982 and accounts for four fifths of it at the 1983 peak, with finance, insurance, and real estate half its size. In 1992–1993 the CPS-based series jump while ours does not: the March CPS replaced its paper questionnaire with computer-assisted interviewing and raised its income recording limits for income year 1993, lifting recorded top incomes and measured inequality with them (Ryscavage 1995; Burkhauser et al. 2012). In 2020–2022 the mean absolute year-over-year change in the top 1% share roughly triples relative to its pre-2010 level, which reflects pandemic-era capital-gains realizations that never enter workplace earnings.

Table 6: Correlation of the U.S. Theil index with published inequality series

| Published series | Levels | First differences |
|---|---|---|
| *A. Top-income shares* | | |
| Top 1% pretax income share (AS) | 0.89 | 0.10 |
| Top 1% pretax income plus transfers (AS) | 0.87 | 0.11 |
| Top 1% fiscal income, including capital gains (PS) | 0.91 | 0.16 |
| *B. Distribution-wide* | | |
| CPS Gini: post-transfer, equivalence-adjusted | 0.91 | 0.09 |
| CPS Theil: post-transfer, equivalence-adjusted | 0.89 | 0.10 |
| BFJL Gini: post-transfer, household | 0.87 | 0.14 |
| BFJL Gini: pretransfer, tax-unit | 0.88 | 0.22 |
| *C. Bottom-99 percent* | | |
| BFJL bottom-99 Gini: post-transfer, household | 0.88 | 0.05 |
| BFJL bottom-99 Gini: pretransfer, tax-unit | 0.89 | 0.17 |
| *D. Correlations among the published series* | | |
| CPS Gini vs. BFJL Gini (both post-transfer, household-based) | 1.00 | 0.92 |
| CPS Gini vs. BFJL Gini: pretransfer, tax-unit | 0.99 | 0.66 |
| Top 1% pretax share (AS) vs. CPS Theil | 0.92 | 0.05 |
| Top 1% pretax share (AS) vs. CPS Gini | 0.92 | 0.09 |
| Top 1% pretax share (AS) vs. BFJL Gini: pretransfer, tax-unit | 0.86 | −0.21 |
| Top 1% pretax share (AS) vs. BFJL bottom-99 Gini: pretransfer, tax-unit | 0.81 | −0.33 |

**Note:** Entries in Panels A–C are Pearson correlations between the Theil index constructed in this paper and the indicated published series; entries in Panel D are correlations among the published series themselves, and do not involve our index. First differences are annual changes. All series measure pretax income. The AS, PS, and CPS series cover 1969–2022 ($n = 54$); the BFJL series cover 1969–2006 ($n = 38$). AS = Auten and Splinter (2024); PS = Piketty and Saez (2003), as reproduced in the AS data files; CPS = U.S. Census Bureau, Table A-5, computed from the March CPS; BFJL = Table C1 of Burkhauser et al. (2012), computed from internal March CPS data. The CPS series measure post-cash-transfer money income, excluding capital gains, equivalence-adjusted with the Census three-parameter scale. The BFJL post-transfer household series measures household-size-adjusted (square-root scale) post-cash-transfer income, excluding capital gains, among persons; the BFJL pretransfer tax-unit series measures non-size-adjusted pretransfer income from taxable sources, excluding capital gains, among tax units. The BFJL bottom-99 series recompute each measure after excluding the top 1% of the respective distribution. Neither CPS-based source adjusts for the 1992–1993 CPS redesign (the so-called "blip").

Figure 8: The U.S. Theil index and other inequality series




**Note:** The thick purple series is the Theil index constructed in this paper. The black series are published series: the top 1% pretax income share (AS); the top 1% fiscal income share, including capital gains (PS); the CPS Theil (post-transfer, equivalence-adjusted); and the BFJL Gini (post-transfer, household). Each series is standardized to mean zero and unit variance over its own available sample, 1969–2022 for all series except the BFJL Gini, which ends in 2006. AS = Auten and Splinter (2024); PS = Piketty and Saez (2003), as reproduced in the AS data files; CPS = U.S. Census Bureau, Table A-5, computed from the March CPS; BFJL = Table C1 of Burkhauser et al. (2012), computed from internal March CPS data. Neither CPS nor BFJL adjusts for the blip arising from the 1992–1993 questionnaire redesign.

## Comparison with existing state inequality series

At the state level, historical inequality series are scarce. The only published long-run benchmark is the Frank et al. (2015) top-income-share panel, which applies the Piketty and Saez (2003) fiscal-income method—tax-return income including realized capital gains—to state-level IRS data. The recent exchange between Auten and Splinter (2024) and Piketty, Saez, and Zucman bears on this series: the AS corrections imply a lower level and a flatter rise for fiscal-income top shares, and to the best of our knowledge, whether or how the state series should be adjusted remains an open question. Table A1 summarizes the state-by-state correlations with the FSP top 1% share; the appendix reports all fifty-one. The median levels correlation is 0.36 (interquartile range 0.09–0.60), positive in four fifths of states, and in 49 of 51 states both series rise over the sample; the median first-differences correlation is −0.05.

Tracking is strongest where the two measures capture the same income concept. Our index records place-of-work earnings; the top-share series record resident fiscal income. The two coincide

in earnings-driven states, where top incomes accrue to the "working rich" (Piketty and Saez 2003) in industries such as finance in New York and technology in California, where bonuses and vested stock are taxed as wage income and produced in the state where their earners file (Bakija et al. 2012). Correlation drops where either condition breaks. In Nevada and Florida, resident top incomes are dominated by retirement income that never enters workplace earnings. Delaware's recorded top incomes reflect ownership structure; Wyoming has a small population and no income tax that attracts "millionaire migrants" (Young et al. 2016). In the District of Columbia the income concept holds but the geography fails: the earnings are produced in the District while the top filers reside in Maryland and Virginia. Consistent with this reading, the earnings-driven states lead not only in levels (0.93 in New York, 0.84 in California) but in first differences (0.59 and 0.71, against the median of −0.05). The pandemic years drag that median: when 2020–2022 is excluded, the median within-state first-differences correlation is +0.03.

Table 7: State-level correlation of the Theil index with the FSP top 1% share

| | Levels | First differences |
|---|---|---|
| Total Theil | 0.36 | −0.05 |
| Between-county component | 0.68 | 0.15 |
| Within-county component | 0.28 | −0.08 |

**Note:** We report the median across states of the within-state Pearson correlation between our Theil series and the Frank–Sommeiller–Price (FSP) top 1% income share (Frank et al. 2015), 1969–2022 ($n = 54$ years per state). First differences are annual changes.

## Usage Notes

The Stata do-files used to generate the data, along with the finalized dataset, will be made available on the UTIP website (http://www.utip.lbj.utexas.edu/datasets.html) and on the Harvard Dataverse.

## Acknowledgements

The authors thank Jeffrey L. Newman and the staff of the Bureau of Economic Analysis, U.S. Department of Commerce, for their help in accessing archived data.

## References


Auten, Gerald, and David Splinter. 2024. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." Data accompanying the article, available from David Splinter's website, *Journal of Political Economy* 132 (7): 2179–2227. https://doi.org/10.1086/728741. https://davidsplinter.com/AutenSplinter-IncomeIneq.xlsx.

Bakija, Jon, Adam Cole, and Bradley T. Heim. 2012. "Jobs and Income Growth of Top Earners and the Causes of Changing Income Inequality: Evidence from U.S. Tax Return Data." Working paper, Williams College.

Bourguignon, François. 1979. "Decomposable Income Inequality Measures." *Econometrica* 47 (4): 901–920.

Bureau of Economic Analysis. 2024. *State Personal Income and Employment: Concepts, Data Sources, and Statistical Methods.* Technical report. U.S. Department of Commerce. https://www.bea.gov/system/files/methodologies/SPI-Methodology.pdf.

Burkhauser, Richard V, Shuaizhang Feng, Stephen P Jenkins, and Jeff Larrimore. 2011. "Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring." *The Journal of Economic Inequality* 9:393–415.

Burkhauser, Richard V., Shuaizhang Feng, Stephen P. Jenkins, and Jeff Larrimore. 2012. "Recent Trends in Top Income Shares in the United States: Reconciling Estimates from March CPS and IRS Tax Return Data." *The Review of Economics and Statistics* 94 (2): 371–388. https://doi.org/10.1162/REST_a_00200.

Conceição, Pedro, James K Galbraith, and Peter Bradford. 2001. "The Theil index in sequences of nested and hierarchic grouping structures: Implications for the measurement of inequality through time, with data aggregated at different levels of industrial classification." *Eastern Economic Journal* 27 (4): 491–514.

Eckert, Fabian, Teresa C. Fort, Peter K. Schott, and Natalie J. Yang. 2021. *Imputing Missing Values in the US Census Bureau's County Business Patterns.* Technical report. National Bureau of Economic Research.

Eckert, Fabian, Ka-leung Lam, Atif Mian, Karsten Muller, Rafael Schwalb, and Amir Sufi. 2022. *The Early County Business Pattern Files: 1946-1974.* Technical report. Princeton Mimeo.

Economic Analysis, Bureau of. 2024. *Local Area Personal Income and Employment: Concepts and Methods.* Technical report. Accessed June 29, 2025. U.S. Department of Commerce. https://www.bea.gov/sites/default/files/methodologies/BEA-Local-Area-Personal-Income-and-Employment-Concepts-and-Methods.pdf.

Fort, Teresa, and Shawn Klimek. 2016. "The effects of industry classification changes on us employment composition." *Tuck School at Dartmouth.*

Frank, Mark W., Estelle Sommeiller, Mark Price, and Emmanuel Saez. 2015. *Frank-Sommeiller-Price Series for Top Income Shares by US States since 1917.* Methodological Note. Accessed August 7, 2026. World Top Incomes Database. https://profiles.shsu.edu/eco_mwf/usstatesWTID.pdf.

Gabaix, Xavier. 2009. "Power Laws in Economics and Finance." *Annual Review of Economics* 1:255–294.

Galbraith, James K. 2022. "A comparison of major world inequality data sets." In *The Routledge Handbook of Comparative Economic Systems,* 195–213. Routledge.

Galbraith, James K, Béatrice Halbach, Aleksandra Malinowska, Amin Shams, and Wenjie Zhang. 2014. "UTIP global inequality data sets 1963-2008: Updates, revisions and quality checks."

Galbraith, James K, and J Travis Hale. 2009. *The Evolution of Economic Inequality in the United States, 1969-2007.* Technical report. UTIP Working Paper.

Jenkins, S. P., R. V. Burkhauser, S. Feng, and J. Larrimore. 2011. "Measuring inequality using censored data: a multiple-imputation approach to estimation and inference." *Journal of the Royal Statistical Society: Series A* 174:63–81.

Kanbur, Ravi. 2024. *Shannon-Theil-Rawls: Information Theory, Inequality and the Veil of Ignorance.* Technical report. ECINEQ, Society for the Study of Economic Inequality.

Kim, Youngsoon, Joongyang Park, and Ae-Jin Ju. 2024. "New approach to measuring income inequality." *Heliyon* 10 (4).

Larrimore, Jeff, Richard V Burkhauser, Gerald Auten, and Philip Armour. 2021. "Recent trends in US income distributions in tax record data using more comprehensive measures of income including real accrued capital gains." *Journal of Political Economy* 129 (5): 1319–1360.

Lustig, Nora. 2020. *The "Missing Rich" in Household Surveys: Causes and Correction Approaches.* ECINEQ Working Paper 2020-520. Society for the Study of Economic Inequality (ECINEQ). https://www.ecineq.org/milano/WP/ECINEQ2020-520.pdf.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *The Quarterly Journal of Economics* 118 (1): 1–39.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2018. "Distributional National Accounts: Methods and Estimates for the United States." *The Quarterly Journal of Economics* 133 (2): 553–609. https://doi.org/10.1093/qje/qjx043.

Ryscavage, Paul. 1995. "A Surge in Growing Income Inequality?" *Monthly Labor Review* 118 (8): 51–61.

Shorrocks, Anthony F. 1980. "The class of additively decomposable inequality measures." *Econometrica: Journal of the Econometric Society,* 613–625.

Theil, Henri. 1967. "Economics and information theory."

———. 1990. "This Week's Citation Classic®." *Economic Forecasts and Policy.*

Young, Cristobal, Charles Varner, Ithai Z. Lurie, and Richard Prisinzano. 2016. "Millionaire Migration and Taxation of the Elite: Evidence from Administrative Data." *American Sociological Review* 81 (3): 421–446. https://doi.org/10.1177/0003122416639625.

Yuskavage, Robert E. 2007. "Converting historical industry time series data from SIC to NAICS." In *The Federal Committee on Statistical Methodology 2007 Research Conference. 5-7 November 2007.* Washington, DC US Department of Commerce, Bureau of Economic Analysis.

# Appendix

# A Derivation of the Theil decomposition for county–sector cells

The derivation follows the logic of Theil's (1967) grouped decomposition, adapted to county–sector cells.

## A.1 Notation and the equality benchmark

Let $g = 1, \ldots, G$ index counties, and let $i \in S_g$ index industrial sectors within county $g$, where $S_g$ denotes the set of sectors included for county $g$. Define

$$
\begin{aligned}
y_i &:= \text{earnings share of sector } i \text{ in county } g \text{ relative to total earnings in the parent economy,} \\
Y_g &:= \sum_{i \in S_g} y_i = \text{earnings share of county } g \text{ relative to total earnings in the parent economy,} \\
n_i &:= \text{employment in sector } i \text{ in county } g, \\
N_g &:= \sum_{i \in S_g} n_i = \text{total employment in county } g, \\
N &:= \sum_{g=1}^{G} N_g = \text{total employment in the parent economy.}
\end{aligned}
$$

By construction,

$$
\sum_{g=1}^{G} \sum_{i \in S_g} y_i = \sum_{g=1}^{G} Y_g = 1, \qquad \sum_{g=1}^{G} \sum_{i \in S_g} \frac{n_i}{N} = \sum_{g=1}^{G} \frac{N_g}{N} = 1.
$$

Each $S_g$ is restricted to county–sector cells with positive employment, so that $n_i/N > 0$ for every included cell. Cells with both zero employment and zero earnings are omitted, and the data are assumed not to contain cells with positive earnings but zero employment:

$$
y_i > 0 \quad \Longrightarrow \quad n_i > 0.
$$

For the included counties, assume that $N_g > 0$ and $Y_g > 0$.

In general, the index compares an observed distribution $\mathbf{y} = (y_i)$ with a benchmark distribution $\mathbf{q} = (q_i)$:

$$
T^* = D_{\mathrm{KL}}(\mathbf{y} \| \mathbf{q}) = \sum_{g=1}^{G} \sum_{i \in S_g} y_i \ln \left( \frac{y_i}{q_i} \right).
$$

For a cell with zero earnings, its contribution is set equal to zero using the standard convention

$$0 \ln\left(\frac{0}{q_i}\right) = 0.$$

In the present setting, individual workers are grouped into county–sector cells, which are in turn grouped into counties, as illustrated in Figures 1 and 2. If every worker in the parent economy earned the same amount, each worker would receive $1/N$ of total earnings. The $n_i$ workers in cell $i$ would therefore jointly receive the earnings share

$$n_i\left(\frac{1}{N}\right) = \frac{n_i}{N}.$$

The equality benchmark is thus the employment-share distribution:

$$q_i = \frac{n_i}{N}, \qquad \text{so that equality requires} \qquad y_i = \frac{n_i}{N} \quad \text{for all } i.$$

## A.2 The county–sector Theil index and its decomposition

Substituting the employment-share benchmark $q_i = n_i/N$ into the general expression gives the county–sector Theil index

$$T^* = \sum_{g=1}^{G} \sum_{i \in S_g} y_i \ln\left(\frac{y_i}{n_i/N}\right). \tag{A.1}$$

This index has the following additive decomposition:

$$\sum_{g=1}^{G} \sum_{i \in S_g} y_i \ln\left(\frac{y_i}{n_i/N}\right) = \underbrace{\sum_{g=1}^{G} Y_g \ln\left(\frac{Y_g}{N_g/N}\right)}_{\text{Between-county inequality}} + \underbrace{\sum_{g=1}^{G} Y_g \left\{ \sum_{i \in S_g} \frac{y_i}{Y_g} \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right) \right\}}_{\text{Within-county inequality}}. \tag{A.2}$$

The between-county component compares each county's earnings share, $Y_g$, with its employment share, $N_g/N$. The within-county component compares the distribution of earnings across sectors within each county, $y_i/Y_g$, with the corresponding distribution of employment, $n_i/N_g$. Each county-specific within-county index is weighted by the county's share of total earnings, $Y_g$.

## A.3 Derivation of the decomposition

To derive the decomposition, consider any sector $i \in S_g$. The ratio inside the logarithm can be factored as

$$\frac{y_i}{n_i/N} = \left(\frac{Y_g}{N_g/N}\right)\left(\frac{y_i/Y_g}{n_i/N_g}\right).$$

Indeed,

$$\left(\frac{Y_g}{N_g/N}\right)\left(\frac{y_i/Y_g}{n_i/N_g}\right) = \left(\frac{Y_g N}{N_g}\right)\left(\frac{y_i N_g}{Y_g n_i}\right) = \frac{y_i N}{n_i} = \frac{y_i}{n_i/N}.$$

Taking logarithms gives

$$\ln\left(\frac{y_i}{n_i/N}\right) = \ln\left(\frac{Y_g}{N_g/N}\right) + \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right).$$

Substituting this identity into (A.1) yields

$$\begin{aligned} T^* &= \sum_{g=1}^{G}\sum_{i\in S_g} y_i\left[\ln\left(\frac{Y_g}{N_g/N}\right) + \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right)\right] \\ &= \sum_{g=1}^{G}\sum_{i\in S_g} y_i \ln\left(\frac{Y_g}{N_g/N}\right) + \sum_{g=1}^{G}\sum_{i\in S_g} y_i \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right). \end{aligned}$$

For the first term, the logarithmic expression does not vary across sectors within county $g$. Therefore,

$$\begin{aligned} \sum_{g=1}^{G}\sum_{i\in S_g} y_i \ln\left(\frac{Y_g}{N_g/N}\right) &= \sum_{g=1}^{G}\left(\sum_{i\in S_g} y_i\right)\ln\left(\frac{Y_g}{N_g/N}\right) \\ &= \sum_{g=1}^{G} Y_g \ln\left(\frac{Y_g}{N_g/N}\right). \end{aligned}$$

This is the between-county component.

For the second term, use $y_i = Y_g(y_i/Y_g)$:

$$\sum_{g=1}^{G}\sum_{i\in S_g} y_i \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right) = \sum_{g=1}^{G} Y_g\left\{\sum_{i\in S_g}\frac{y_i}{Y_g}\ln\left(\frac{y_i/Y_g}{n_i/N_g}\right)\right\}.$$

This is the within-county component. Combining the two expressions establishes the decomposition in (A.2).

### A.4 Interpretation

The first term,

$$\sum_{g=1}^{G} Y_g \ln\left(\frac{Y_g}{N_g/N}\right),$$

measures between-county inequality. It compares each county's share of total earnings with its share of total employment. The county-specific between-county element is

$$B_g \equiv Y_g \ln\left(\frac{Y_g}{N_g/N}\right).$$

Because $Y_g > 0$, the sign of $B_g$ is determined by the ratio $Y_g/(N_g/N)$. In particular,

$$B_g \begin{cases} > 0, & \text{if } Y_g > N_g/N, \\ = 0, & \text{if } Y_g = N_g/N, \\ < 0, & \text{if } Y_g < N_g/N. \end{cases}$$

Thus, a county's between-county element is positive when its earnings share exceeds its employment share and negative when its earnings share falls below its employment share. Although individual county elements may be negative, their sum is nonnegative. The aggregate between-county component is zero if and only if average earnings per worker are equal across counties.

The second term,

$$\sum_{g=1}^{G} Y_g \left\{ \sum_{i \in S_g} \frac{y_i}{Y_g} \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right) \right\},$$

measures within-county inequality across sectors. Within county $g$, define the unweighted sectoral Theil index as

$$T_g^W \equiv \sum_{i \in S_g} \frac{y_i}{Y_g} \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right).$$

The contribution of county $g$ to the aggregate within-county component is therefore

$$W_g \equiv Y_g T_g^W = Y_g \left\{ \sum_{i \in S_g} \frac{y_i}{Y_g} \ln\left(\frac{y_i/Y_g}{n_i/N_g}\right) \right\}.$$

Each $W_g$ is nonnegative because $T_g^W$ is a Kullback–Leibler divergence between the sectoral earnings-share and employment-share distributions within county $g$. It equals zero if and only if

$$\frac{y_i}{Y_g} = \frac{n_i}{N_g} \qquad \text{for every } i \in S_g,$$

which means that average earnings per worker are equal across sectors within that county.

Because $W_g$ equals the county's share of national earnings (or state earnings, depending on the parent economy) multiplied by its within-county sectoral Theil index, a large contribution may reflect a large earnings share, a large disparity between sectoral earnings and employment shares within the county, or both. Rankings based on $W_g$ therefore rank counties by their contributions to the aggregate within-county component; they should not be interpreted as rankings of unweighted local inequality.

### A.5 Nonnegativity and equality conditions

Both components in equation (A.2) are nonnegative because they are KL divergences, with the county-specific within-county divergences weighted by $Y_g > 0$.

The between-county component is zero if and only if

$$Y_g = \frac{N_g}{N} \qquad \text{for every } g,$$

meaning that average earnings per worker are equal across counties. The within-county component is zero if and only if

$$\frac{y_i}{Y_g} = \frac{n_i}{N_g} \qquad \text{for every } i \in S_g \text{ and every } g,$$

meaning that average earnings per worker are equal across sectors within each county.

Because the total index is the sum of the between- and within-county components, and the within-county component is nonnegative, the between-county component provides a lower bound for the total index:

$$T^* \geq \sum_{g=1}^{G} Y_g \ln\left(\frac{Y_g}{N_g/N}\right) \geq 0.$$

The first inequality holds with equality if and only if the within-county component is zero, while the second holds with equality if and only if the between-county component is zero. Therefore, the total index is zero if and only if both components are zero. Equivalently,

$$T^* = 0 \quad \Longleftrightarrow \quad y_i = \frac{n_i}{N} \qquad \text{for every included county–sector cell.}$$

### A.6 Bounds of $T^*$

For purposes of establishing the bounds of the index, it is convenient to suppress the county grouping and re-index all included county–sector cells by $i = 1, \ldots, I$. Each included cell has positive employment. Let $y_i$ denote cell $i$'s share of total earnings and define its employment share as

$$q_i \equiv \frac{n_i}{N},$$

where $n_i$ is employment in cell $i$ and

$$N = \sum_{i=1}^{I} n_i.$$

Thus,

$$y_i \geq 0, \qquad q_i > 0, \qquad \sum_{i=1}^{I} y_i = \sum_{i=1}^{I} q_i = 1.$$

Holding the employment-share distribution $\mathbf{q}$ fixed and allowing $\mathbf{y}$ to vary over all earnings-share distributions, the inequality measure is

$$T^* = \sum_{i=1}^{I} y_i \ln\left(\frac{y_i}{q_i}\right) = \sum_{i=1}^{I} y_i \ln\left(\frac{y_i}{n_i/N}\right) = D_{\mathrm{KL}}(\mathbf{y}\|\mathbf{q}),$$

where the standard convention

$$0 \ln\left(\frac{0}{q_i}\right) = 0$$

is used.

### A.6.1 Minimum

We use the elementary inequality

$$\ln x \leq x - 1,$$

with equality if and only if $x = 1$. For each cell with $y_i > 0$, set $x = q_i/y_i$. Then

$$\begin{aligned} \ln x \leq x - 1 \quad &\Longrightarrow \quad \ln\left(\frac{q_i}{y_i}\right) \leq \frac{q_i}{y_i} - 1 \\ &\Longrightarrow \quad -y_i \ln\left(\frac{q_i}{y_i}\right) \geq y_i - q_i \\ &\Longrightarrow \quad y_i \ln\left(\frac{y_i}{q_i}\right) \geq y_i - q_i. \end{aligned}$$

Summing over all cells yields

$$\begin{aligned} y_i \ln\left(\frac{y_i}{q_i}\right) \geq y_i - q_i & \\ &\Longrightarrow \quad \sum_{i=1}^{I} y_i \ln\left(\frac{y_i}{q_i}\right) \geq \sum_{i=1}^{I} y_i - \sum_{i=1}^{I} q_i \\ &\Longrightarrow \quad T^* \geq 1 - 1 \\ &\Longrightarrow \quad T^* \geq 0. \end{aligned}$$

This is the normalized special case of the log–sum inequality:

$$\begin{aligned} \sum_{i=1}^{I} y_i \ln\left(\frac{y_i}{q_i}\right) \geq \left(\sum_{i=1}^{I} y_i\right) \ln\left(\frac{\sum_{i=1}^{I} y_i}{\sum_{i=1}^{I} q_i}\right) & \\ &\Longrightarrow \quad T^* \geq 1 \ln(1) \\ &\Longrightarrow \quad T^* \geq 0, \end{aligned}$$

because both distributions sum to one.

Equality in $\ln x \leq x - 1$ holds if and only if $x = 1$. Therefore,

$$\begin{aligned} T^* = 0 \quad &\iff \quad \frac{q_i}{y_i} = 1 \quad \text{for every } i \\ &\iff \quad y_i = q_i \quad \text{for every } i \\ &\iff \quad D_{\mathrm{KL}}(\mathbf{y} \| \mathbf{q}) = 0. \end{aligned}$$

Substituting $q_i = n_i/N$ gives

$$\min_{\mathbf{y}} T^* = 0 \quad \iff \quad y_i = \frac{n_i}{N} \quad \text{for every } i.$$

Thus, $T^*$ is minimized when each cell's earnings share equals its employment share. If $y_i = C_i/C$, then

$$\begin{aligned} y_i = \frac{n_i}{N} \quad &\iff \quad \frac{C_i}{C} = \frac{n_i}{N} \\ &\iff \quad \frac{C_i}{n_i} = \frac{C}{N}. \end{aligned}$$

Hence, aggregate earnings are distributed proportionally to employment, and average earnings per worker are identical across all cells.

### A.6.2 Maximum

Let $I$ denote the total number of included county–sector cells, and define

$$q_{\min} \equiv \min_{1 \leq i \leq I} q_i.$$

For a fixed employment-share distribution $\mathbf{q}$, consider

$$T^*(\mathbf{y}) = \sum_{i=1}^{I} y_i \ln \left( \frac{y_i}{q_i} \right),$$

where

$$y_i \geq 0, \qquad \sum_{i=1}^{I} y_i = 1.$$

For each cell, define

$$f_i(y_i) = y_i \ln \left( \frac{y_i}{q_i} \right).$$

For $y_i > 0$,

$$f_i''(y_i) = \frac{1}{y_i}$$
$$\implies \quad f_i''(y_i) > 0.$$

Thus, each $f_i$ is convex in $y_i$, and therefore $T^*$ is convex in $\mathbf{y}$. A convex function on the simplex attains its maximum at a vertex. Each vertex represents an earnings distribution in which one cell receives all earnings:

$$y_k = 1, \qquad y_j = 0 \quad \text{for every } j \neq k.$$

The upper bound and the choice of maximizing vertex can be seen directly by rewriting the index as

$$T^* = \sum_{i=1}^{I} y_i \ln y_i + \sum_{i=1}^{I} y_i \ln\left(\frac{1}{q_i}\right).$$

Because $0 \leq y_i \leq 1$,

$$\sum_{i=1}^{I} y_i \ln y_i \leq 0.$$

Moreover,

$$\begin{aligned} q_i \geq q_{\min} \quad &\implies \quad \frac{1}{q_i} \leq \frac{1}{q_{\min}} \\ &\implies \quad \ln\left(\frac{1}{q_i}\right) \leq \ln\left(\frac{1}{q_{\min}}\right). \end{aligned}$$

It follows that

$$\begin{aligned} T^* &= \sum_{i=1}^{I} y_i \ln y_i + \sum_{i=1}^{I} y_i \ln\left(\frac{1}{q_i}\right) \\ &\implies \quad T^* \leq \sum_{i=1}^{I} y_i \ln\left(\frac{1}{q_i}\right) \\ &\implies \quad T^* \leq \sum_{i=1}^{I} y_i \ln\left(\frac{1}{q_{\min}}\right) \\ &\implies \quad T^* \leq \ln\left(\frac{1}{q_{\min}}\right). \end{aligned}$$

The first inequality holds with equality only when all earnings are concentrated in a single cell, so that

$$\sum_{i=1}^{I} y_i \ln y_i = 0.$$

The second inequality holds with equality when positive earnings shares are assigned only to cells satisfying $q_i = q_{\min}$. Consequently, both inequalities hold with equality simultaneously when all earnings are concentrated in one cell having the smallest employment share.

Specifically, select any cell $k$ such that

$$q_k = q_{\min}$$

and set

$$y_k = 1, \qquad y_j = 0 \quad \text{for every } j \neq k.$$

Using the convention $0 \ln(0/q_j) = 0$,

$$\begin{aligned} T^* &= 1 \ln\left(\frac{1}{q_k}\right) + \sum_{j \neq k} 0 \ln\left(\frac{0}{q_j}\right) \\ \implies \quad T^* &= \ln\left(\frac{1}{q_k}\right) \\ \implies \quad T^* &= \ln\left(\frac{1}{q_{\min}}\right). \end{aligned}$$

Therefore,

$$\max_{\mathbf{y}} T^* = \ln\left(\frac{1}{q_{\min}}\right).$$

Since $q_i = n_i/N$, define

$$n_{\min} \equiv \min_{1 \leq i \leq I} n_i.$$

Because every included cell has positive employment,

$$q_{\min} = \frac{n_{\min}}{N}.$$

It follows that

$$\begin{aligned} \max_{\mathbf{y}} T^* &= \ln\left(\frac{1}{q_{\min}}\right) \\ \implies \quad \max_{\mathbf{y}} T^* &= \ln\left(\frac{N}{n_{\min}}\right). \end{aligned}$$

Hence, for a fixed employment distribution,

$$0 \leq T^* \leq \ln\left(\frac{N}{n_{\min}}\right).$$

The maximum occurs when all earnings are concentrated in one of the cells with the smallest positive employment. If several cells have the same minimum employment, concentrating all earnings

in any one of them produces the same maximum.

### A.7 Extending Theil's (1967) entropy representation to county–sector cells

In Theil's original formulation, $N$ denotes the number of units, which are treated as indivisible. Because each unit counts once, the reference distribution is uniform, assigning the benchmark share $1/N$ to every unit.

In the present formulation, the underlying units are workers grouped into $I$ county–sector cells, where $I$ denotes the number of included cells. Cell $i$ contains $n_i$ workers, so

$$N = \sum_{i=1}^{I} n_i, \qquad q_i = \frac{n_i}{N},$$

where $N$ is total employment in the parent economy. Consequently, the reference distribution is uniform across the underlying workers but generally nonuniform across county–sector cells.

Define Shannon entropy as

$$H(\mathbf{y}) = -\sum_{i=1}^{I} y_i \ln y_i.$$

Expanding the logarithm in $T^*$ gives

$$\begin{aligned} T^* &= \sum_{i=1}^{I} y_i \ln\left(\frac{y_i}{q_i}\right) \\ &\Longrightarrow \quad T^* = \sum_{i=1}^{I} y_i \ln y_i - \sum_{i=1}^{I} y_i \ln q_i \\ &\Longrightarrow \quad T^* = -H(\mathbf{y}) - \sum_{i=1}^{I} y_i \ln q_i. \end{aligned}$$

When the benchmark shares $q_i$ are unequal, the cross-entropy term

$$-\sum_{i=1}^{I} y_i \ln q_i$$

depends on which cells receive the earnings and therefore varies with $\mathbf{y}$. Consequently, $T^*$ cannot generally be written as a fixed constant minus $H(\mathbf{y})$. As shown in the preceding subsection,

$$\min_{\mathbf{y}} T^* = 0 \quad \text{when} \quad y_i = q_i \quad \text{for every } i,$$

whereas

$$\max_{\mathbf{y}} T^* = \ln\left(\frac{1}{q_{\min}}\right), \qquad q_{\min} \equiv \min_{1 \le i \le I} q_i.$$

Thus, the minimum does not generally coincide with maximum Shannon entropy, while the maximum is determined by the smallest benchmark share.

The constant-minus-entropy representation is recovered when the benchmark distribution is uniform. If all cells have equal employment shares, so that $q_i = 1/I$, then

$$-\sum_{i=1}^{I} y_i \ln q_i = \ln I,$$

and therefore

$$T^* = \ln I - H(\mathbf{y}).$$

Because

$$0 \leq H(\mathbf{y}) \leq \ln I,$$

it follows that

$$0 \leq T^* \leq \ln I.$$

The minimum occurs when $y_i = 1/I$ for every cell, and the maximum occurs when all earnings are concentrated in a single cell.

Theil's original result is recovered when the grouping is removed and each worker is entered separately. In that case,

$$I = N, \qquad n_i = 1, \qquad q_i = \frac{1}{N},$$

so

$$T^* = \ln N - H(\mathbf{y}), \qquad 0 \leq T^* \leq \ln N.$$

Thus, the present measure extends Theil's Kullback–Leibler form to county–sector cells containing different numbers of workers. The relevant benchmark at the county–sector level is $q_i = n_i/N$, rather than $1/I$. Under equal earnings per worker, a cell containing $n_i$ workers must receive the benchmark earnings share $n_i/N$.

### A.8 Relation to worker-level inequality

To relate $T^*$ to inequality among workers, let $j = 1, \ldots, n_i$ index workers in county–sector cell $i$, and let $x_{ij}$ denote the earnings of worker $j$ in that cell. Define

$$X_i = \sum_{j=1}^{n_i} x_{ij}, \qquad X = \sum_{i=1}^{I} X_i, \qquad N = \sum_{i=1}^{I} n_i,$$

and

$$y_i = \frac{X_i}{X}, \qquad q_i = \frac{n_i}{N}.$$

For $X_i > 0$, define worker $j$'s conditional earnings share within cell $i$ as

$$p_{ij} = \frac{x_{ij}}{X_i}, \qquad \sum_{j=1}^{n_i} p_{ij} = 1.$$

Cells with $X_i = 0$ have $y_i = 0$ and contribute zero to the decomposition. The following identities hold:

$$\left.\begin{aligned} \frac{x_{ij}}{X} &= y_i p_{ij}, \\ \frac{1}{N} &= q_i \frac{1}{n_i} \end{aligned}\right\} \quad \Longrightarrow \quad \frac{x_{ij}/X}{1/N} = \frac{y_i}{q_i}\frac{p_{ij}}{1/n_i}$$

$$\Longrightarrow \quad \ln\left(\frac{x_{ij}/X}{1/N}\right) = \ln\left(\frac{y_i}{q_i}\right) + \ln\left(\frac{p_{ij}}{1/n_i}\right).$$

Define the Theil index among workers within cell $i$ as

$$T_i = \sum_{j=1}^{n_i} p_{ij} \ln\left(\frac{p_{ij}}{1/n_i}\right).$$

Equivalently, letting $\bar{x}_i = X_i/n_i$,

$$T_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{x_{ij}}{\bar{x}_i} \ln\left(\frac{x_{ij}}{\bar{x}_i}\right).$$

Substitution into the worker-level Theil index gives

$$\begin{aligned} T_{\text{workers}} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} y_i p_{ij}\left[\ln\left(\frac{y_i}{q_i}\right) + \ln\left(\frac{p_{ij}}{1/n_i}\right)\right] \\ &= \sum_{i=1}^{I} y_i \ln\left(\frac{y_i}{q_i}\right) \underbrace{\sum_{j=1}^{n_i} p_{ij}}_{=1} + \sum_{i=1}^{I} y_i \underbrace{\sum_{j=1}^{n_i} p_{ij} \ln\left(\frac{p_{ij}}{1/n_i}\right)}_{=T_i} \\ &= T^* + \sum_{i=1}^{I} y_i T_i. \end{aligned} \tag{A.3}$$

The worker-level interpretation of $T^*$ follows by letting $\bar{x} = \frac{X}{N}$ denote average earnings across all workers:

$$\frac{y_i}{q_i} = \frac{X_i/X}{n_i/N} = \frac{X_i/n_i}{X/N} = \frac{\bar{x}_i}{\bar{x}}$$

$$\implies \quad T^* = \sum_{i=1}^{I} \frac{n_i}{N} \frac{\bar{x}_i}{\bar{x}} \ln\left(\frac{\bar{x}_i}{\bar{x}}\right).$$

To see this directly, assign worker $j$ in cell $i$ the cell's average earnings,

$$\widetilde{x}_{ij} = \bar{x}_i.$$

Then

$$\widetilde{x}_{ij} = \bar{x}_i \quad \implies \quad \frac{\widetilde{x}_{ij}/X}{1/N} = \frac{\bar{x}_i}{\bar{x}}$$

$$\implies \quad T(\widetilde{x}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \frac{\bar{x}_i}{X} \ln\left(\frac{\bar{x}_i}{\bar{x}}\right) = \sum_{i=1}^{I} \frac{n_i}{N} \frac{\bar{x}_i}{\bar{x}} \ln\left(\frac{\bar{x}_i}{\bar{x}}\right) = T^*.$$

Thus, $T^*$ is exactly the inequality between workers across county–sector cells attributable to differences in average earnings across those cells. The unobserved remainder, $\sum_{i=1}^{I} y_i T_i$, is the earnings-share-weighted sum of inequality among workers employed in the same county and sector.

Because each within-cell Theil index is nonnegative,

$$T_i \geq 0 \quad \implies \quad \sum_{i=1}^{I} y_i T_i \geq 0 \quad \implies \quad T_{\text{workers}} \geq T^*.$$

Therefore, $T^*$ is a lower bound on total worker-level inequality. Equality holds if and only if earnings are equal among workers within every county–sector cell with positive total earnings.

Finally, because the preceding decomposition establishes that

$$T^* = T_{\text{between}} + T_{\text{within}},$$

Equation A.3 implies

$$T_{\text{workers}} = \underbrace{T_{\text{between}}}_{\text{inequality across counties}} + \underbrace{T_{\text{within}}}_{\text{inequality across county–sector cells within counties}} + \underbrace{\sum_{i=1}^{I} y_i T_i}_{\text{inequality within county–sector cells}}.$$

# B Placebo Tests of Sectoral Share Stability

We assess the stability of sectoral composition under both the SIC and NAICS regimes using placebo exercises. Within each eligible county–year, $K$ observed sectors are randomly suppressed and reconstructed using nearest-year sectoral shares.

### Imputation Procedure Under the SIC Regime

Across 3,139 county×year observations, the mean number of missing sectors equals 0.39 and the median equals 0. The 90th percentile equals two. Only 1% of county–years have three or more missing sectors. Thus, when imputation is required, it typically involves a single sector.

The imputation procedure allocates the residual between the published county total (`LineCode` = 10) and the sum of observed sectors according to sectoral shares from the nearest year in which the missing sector is observed within that county. Empirically, the selected reference year is almost always adjacent to the missing year.

We evaluate whether using historical shares is reasonable through a placebo exercise calibrated to the data structure. Specifically, within each eligible county×year (defined as having more than 8 out of 11 sectors reported), we randomly suppress one observed sector and reconstruct it using sectoral shares from the nearest year.

The placebo results indicate that the procedure reproduces sector-level values with high accuracy. A regression of true earnings on imputed earnings yields a slope coefficient of 0.997 (robust standard error 0.002) and an $R^2$ of 0.999.

The mean absolute error (MAE) are small. For employment, the MAE equals 173, which corresponds to 1.6% of the mean and 0.14% of the cross-sectional standard deviation. For earnings, the MAE equals approximately \$4.1 million, or 1.7% of the mean and 0.12% of the standard deviation. Relative to the median sector size, the MAE equals 28% for employment and 56% for earnings. Because the distribution of sectoral values is highly skewed, the median represents a relatively small sector. These ratios therefore reflect proportional noise for small sectors rather than meaningful distortion of the overall distribution. Taken together, these results indicate that the imputation approach under the SIC regime is well supported by the data.

### Imputation Procedure Under the NAICS Regime

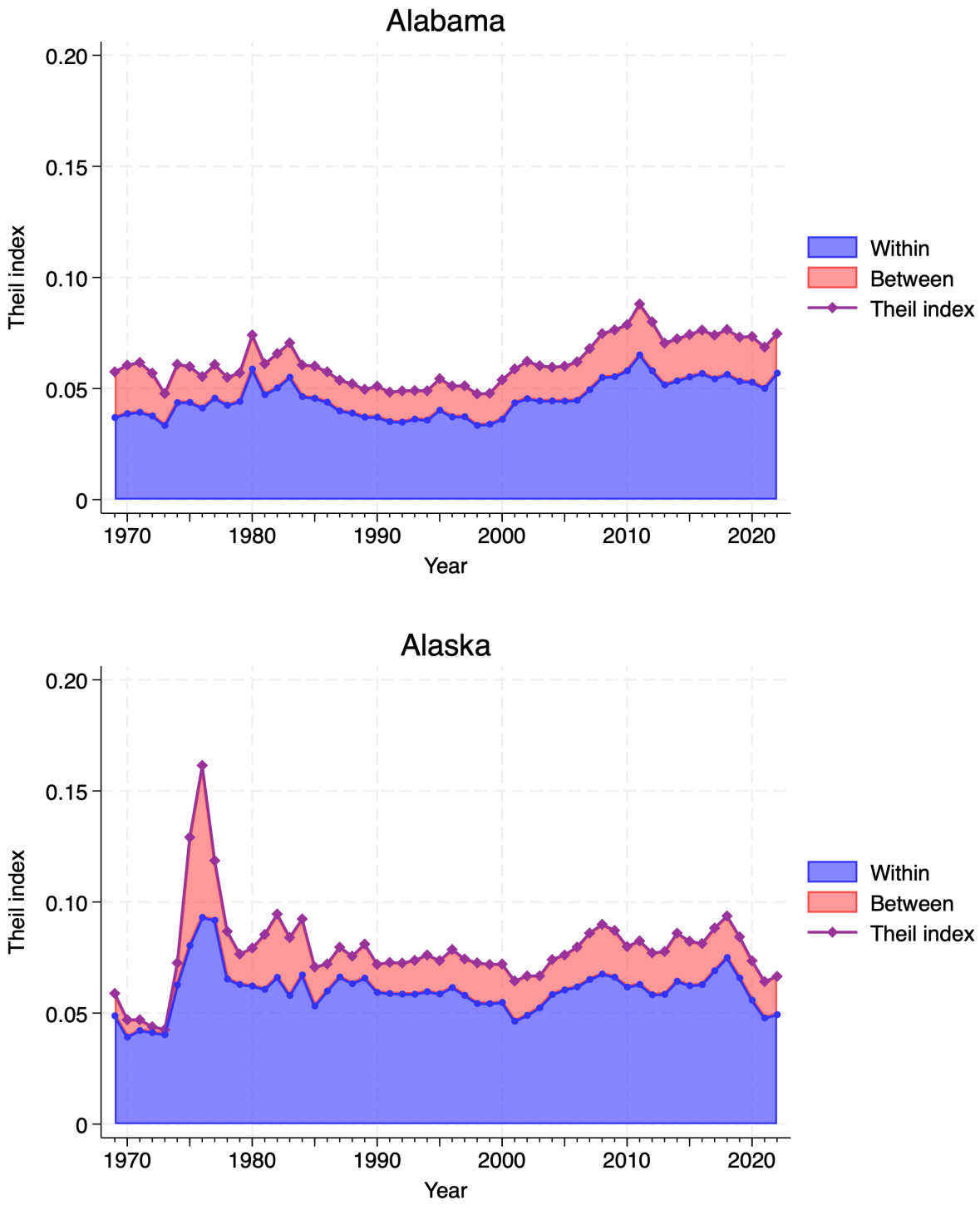
Under the NAICS regime, the mean number of missing sectors per county×year equals five. We therefore calibrate the placebo exercise by randomly suppressing 5 observed sectors within each eligible county×year and reconstructing them using nearest-year sectoral shares.

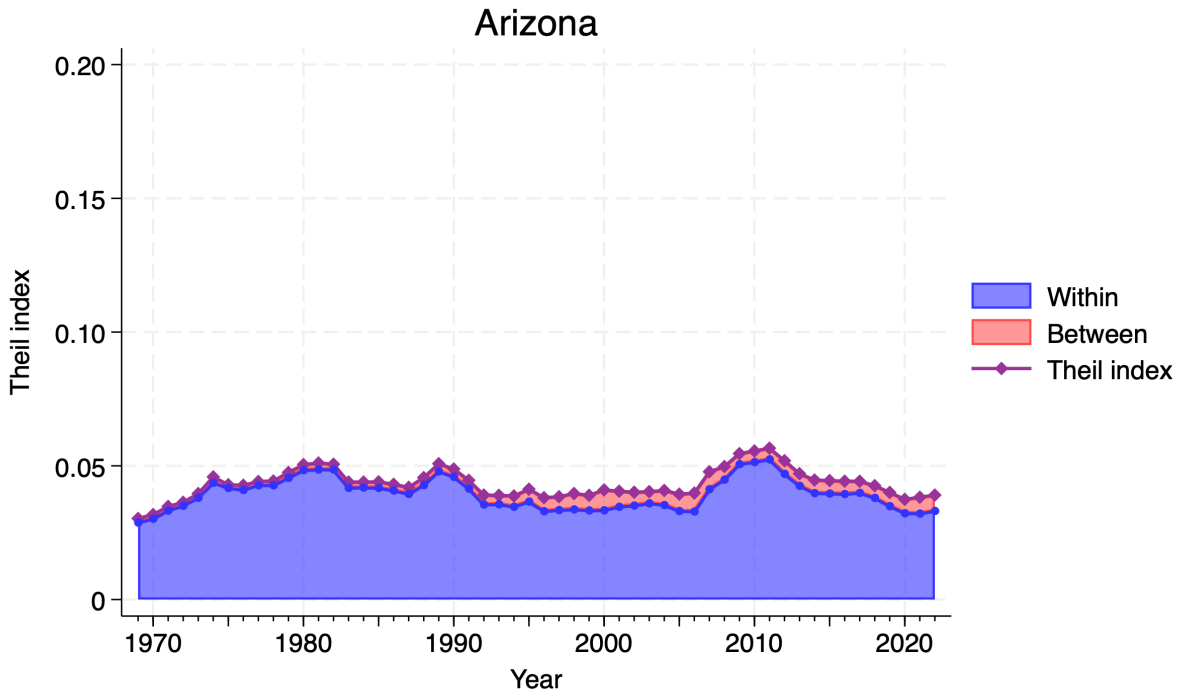
When true values are regressed on imputed values, the slope coefficients equal 0.730 for employment and 0.734 for earnings (robust standard errors 0.017 and 0.019, respectively). The intercepts are not statistically different from zero, and the $R^2$ values equal 0.904 for employment and 0.896 for earnings. The slope below one indicates that missing values are imputed conservatively when five sectors are reconstructed simultaneously.

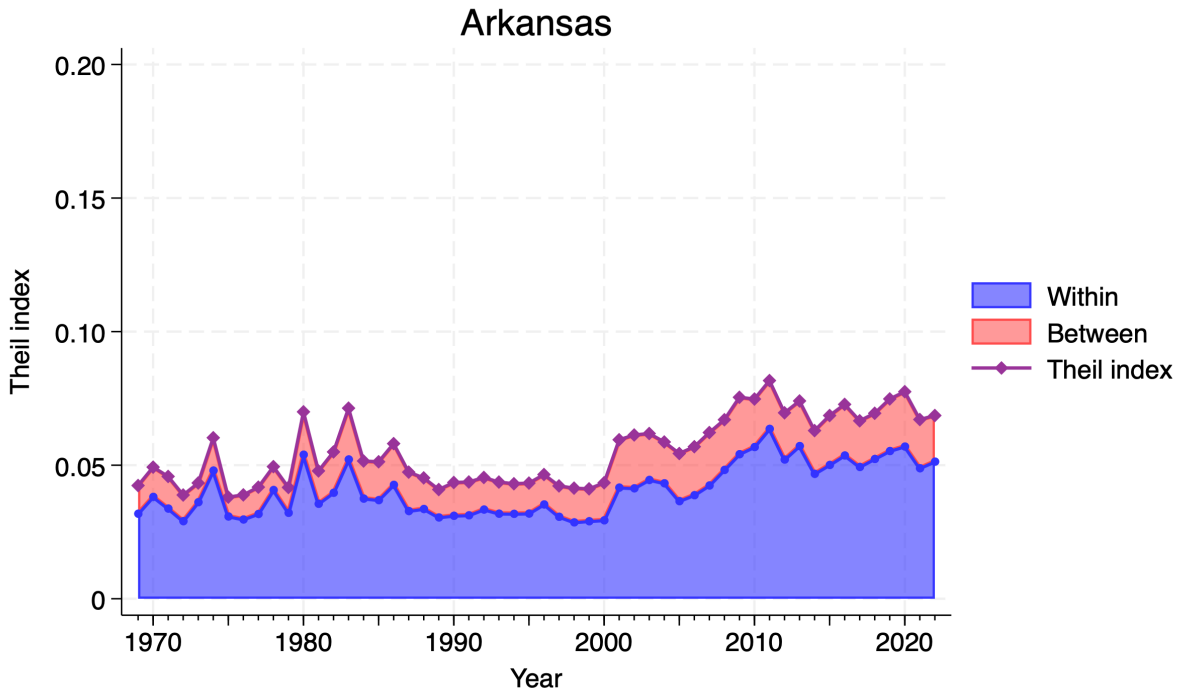
The MAE is moderate relative to dispersion. For employment, the MAE equals 7,989, or 5.6% of the cross-sectional standard deviation and 37% of the mean. For earnings, the MAE equals $462 million, or 5.4% of the standard deviation and 38% of the mean. Relative to the median sector size, the MAE equals 4.63 for employment and 7.45 for earnings. Because sectoral values are highly skewed and the median represents a small sector, these ratios mainly reflect larger proportional errors for small sectors.

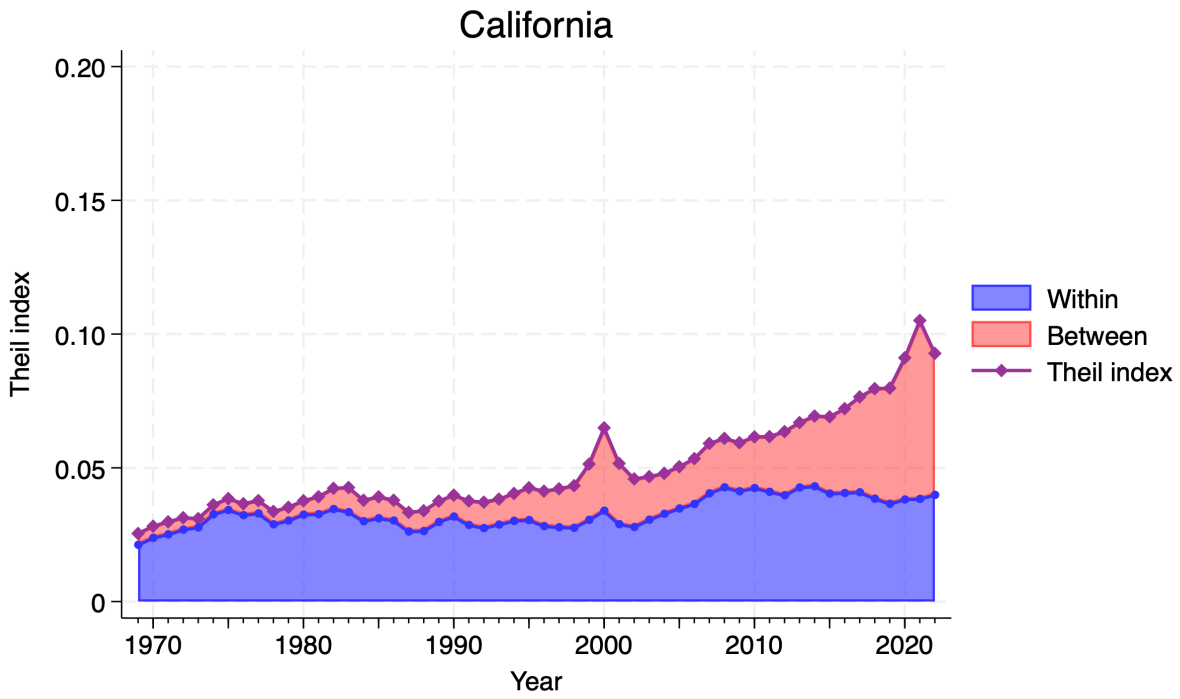
The median absolute percent error equals 36.2% for employment and 36.9% for earnings. Correlations between true and imputed values equal 0.951 for employment and 0.946 for earnings. Overall, the nearest-year share method fits the data well but yields conservative imputations, with lower cross-sectional variation than the true values. This procedure imputes approximately 11% of missing sector cells in the NAICS files, compared with 91% in the SIC files.

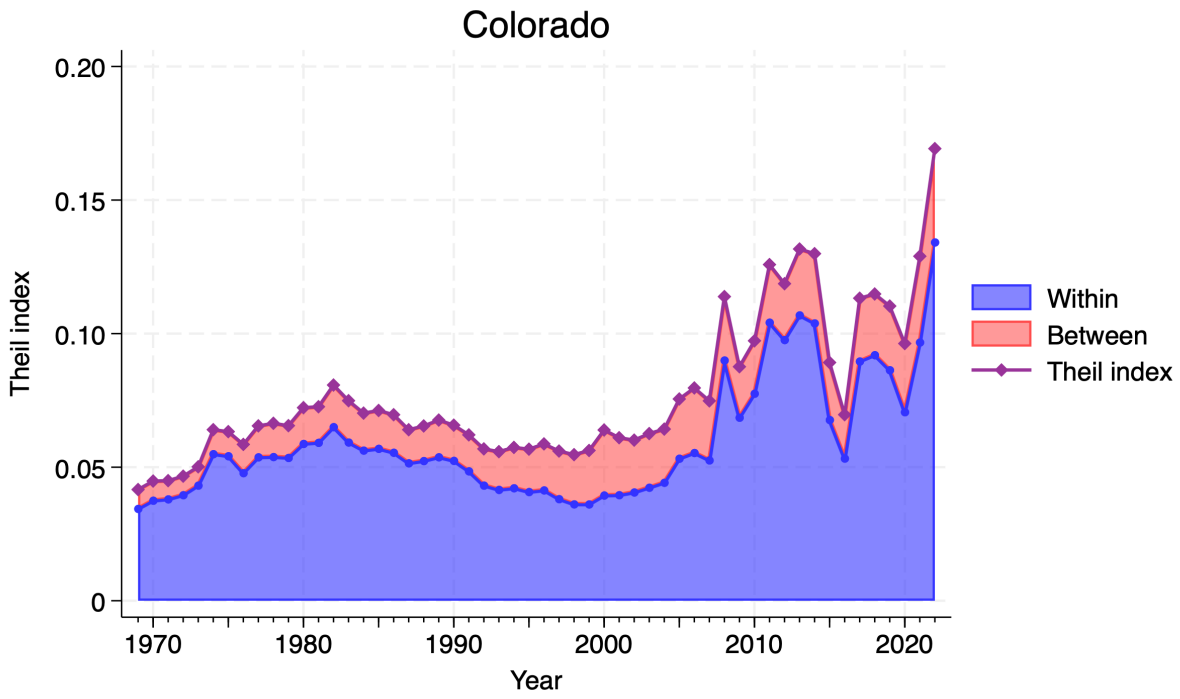
## C State-by-state evolution of inequality decomposed into between and within inequality, 1969–2022

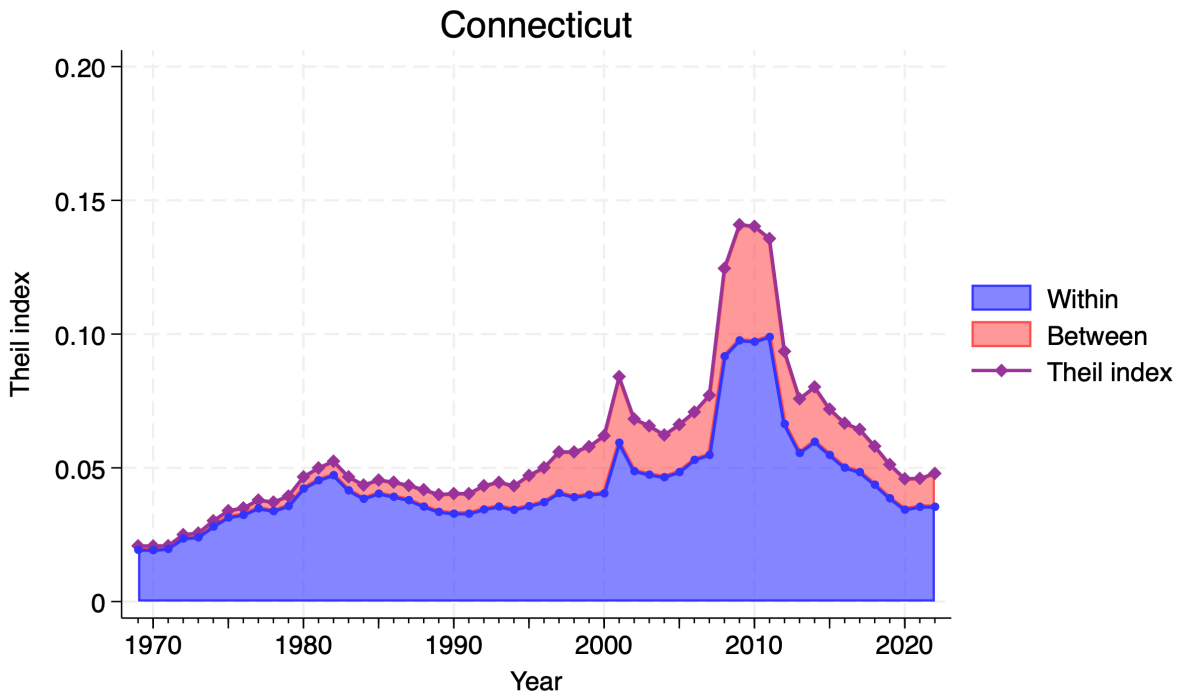


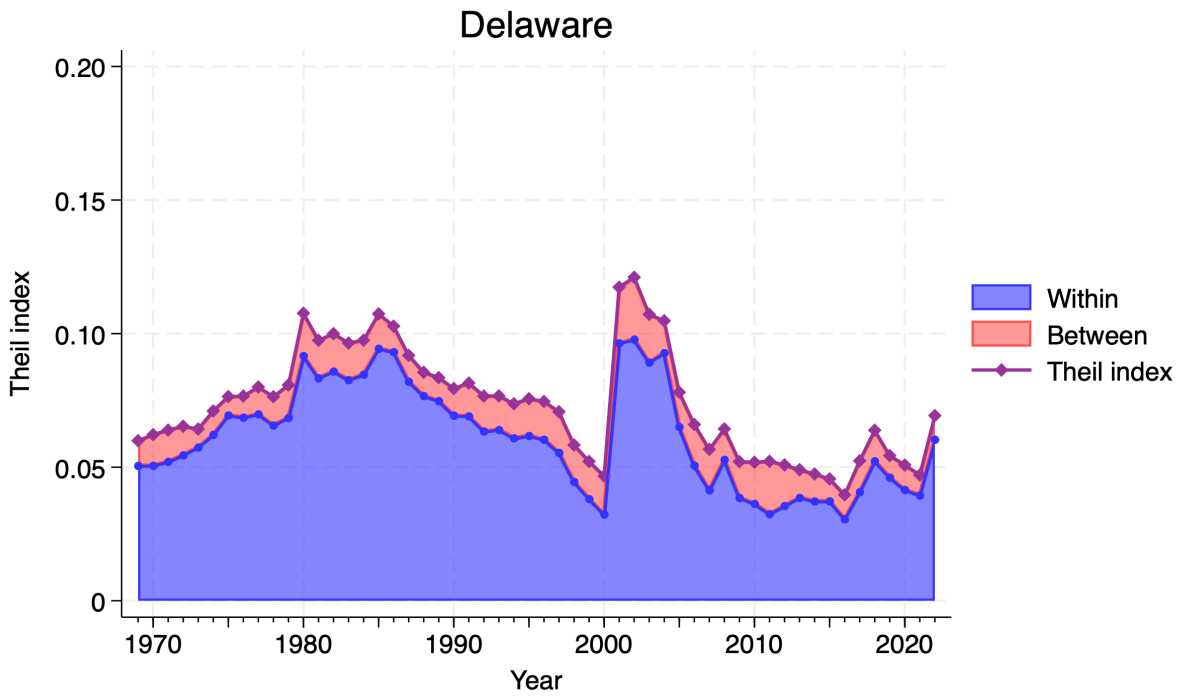

Arizona
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

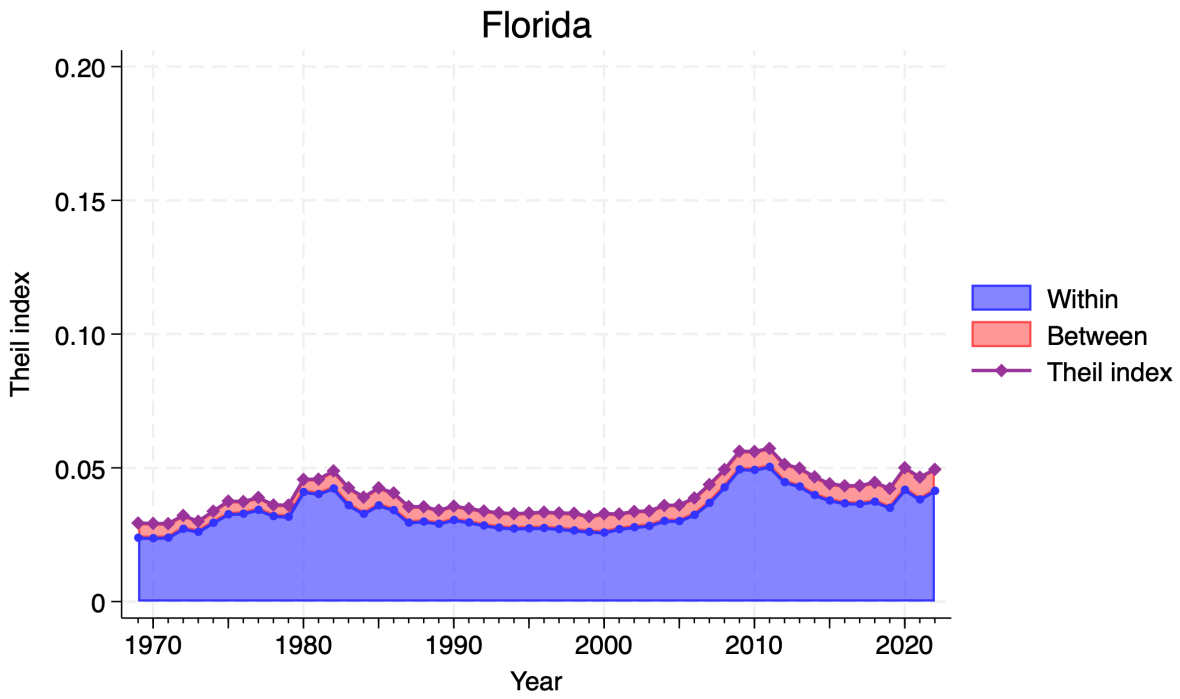

Arkansas
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

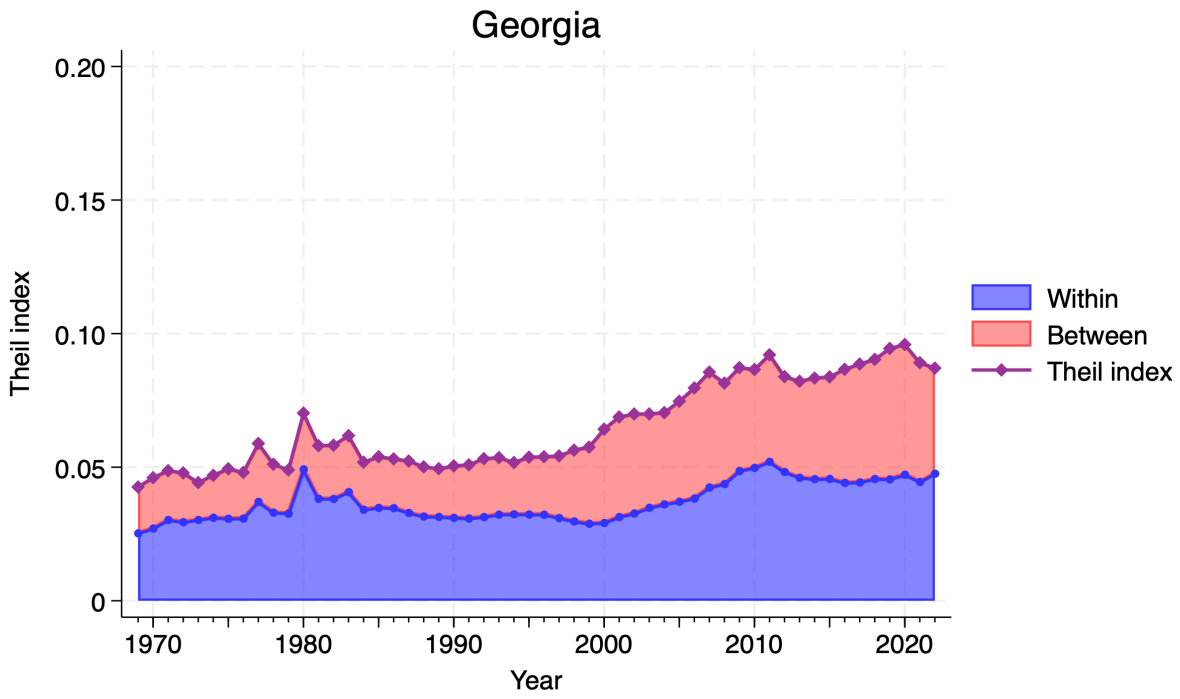

California
Theil index
0.20
0.15
0.10
0.05
0
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

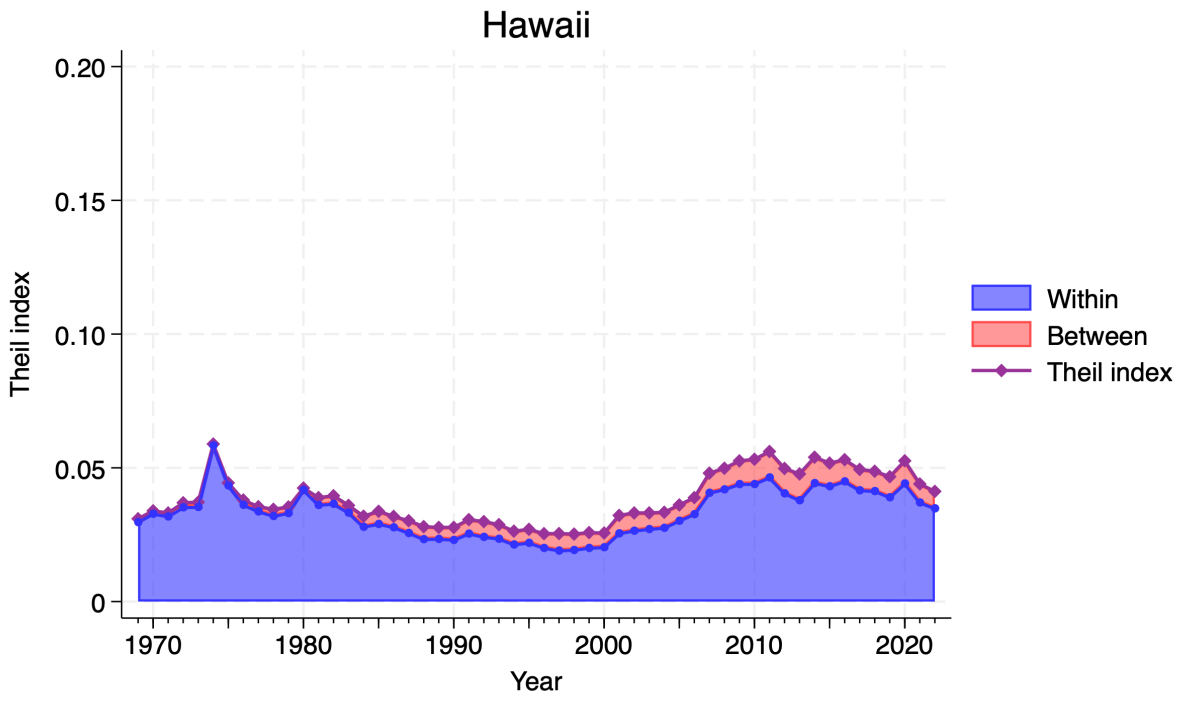

Colorado
Theil index
0.20
0.15
0.10
0.05
0
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

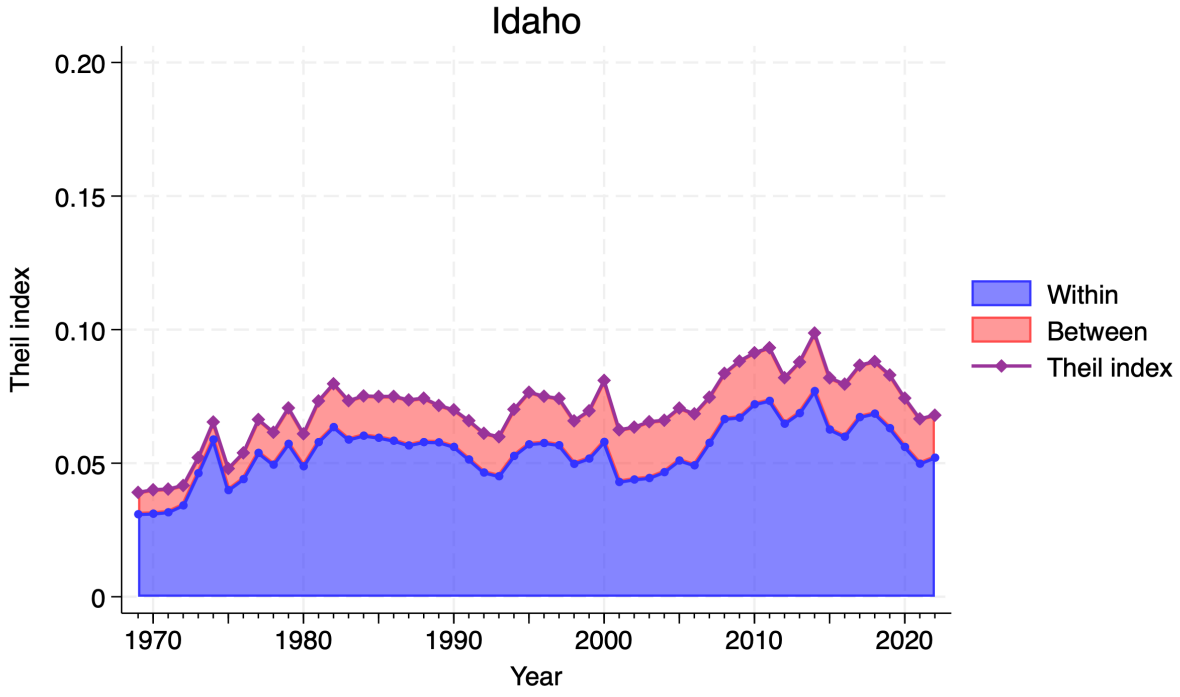

Connecticut
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

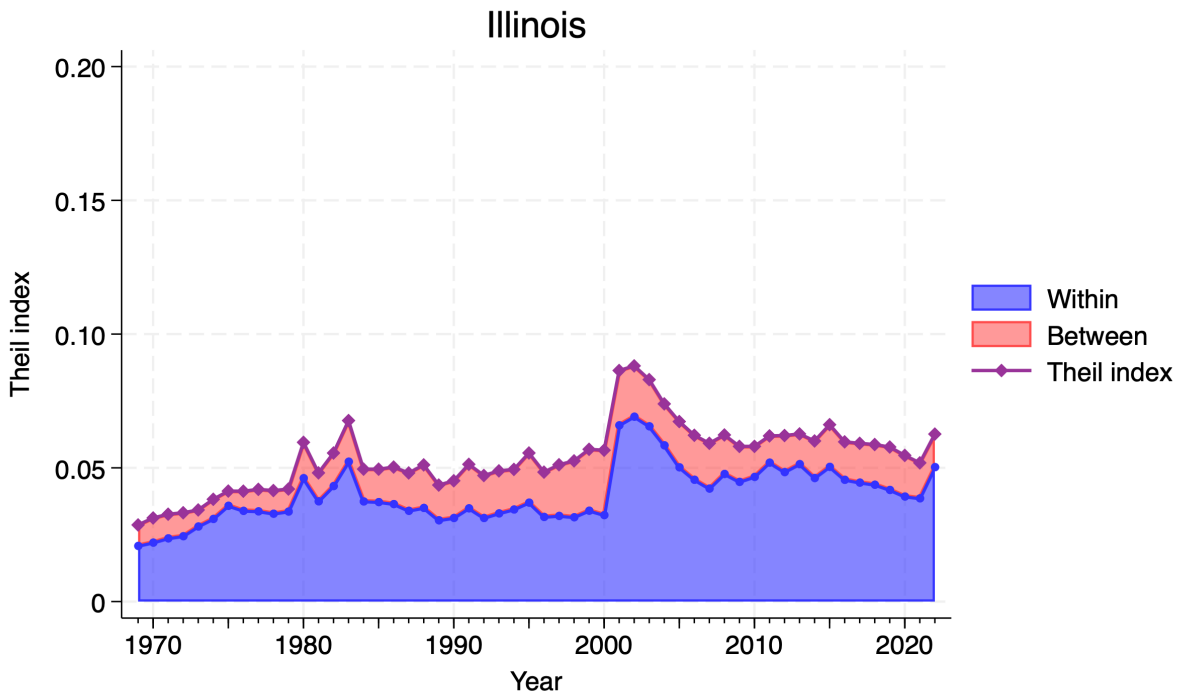

Delaware
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

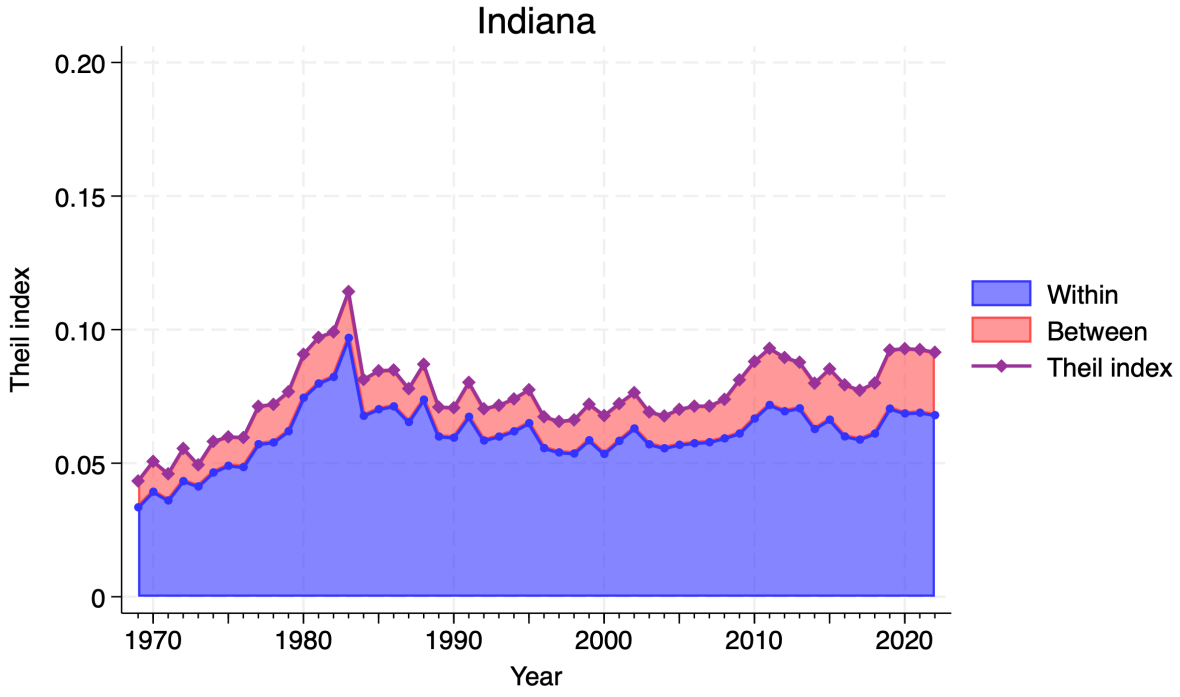

Florida
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

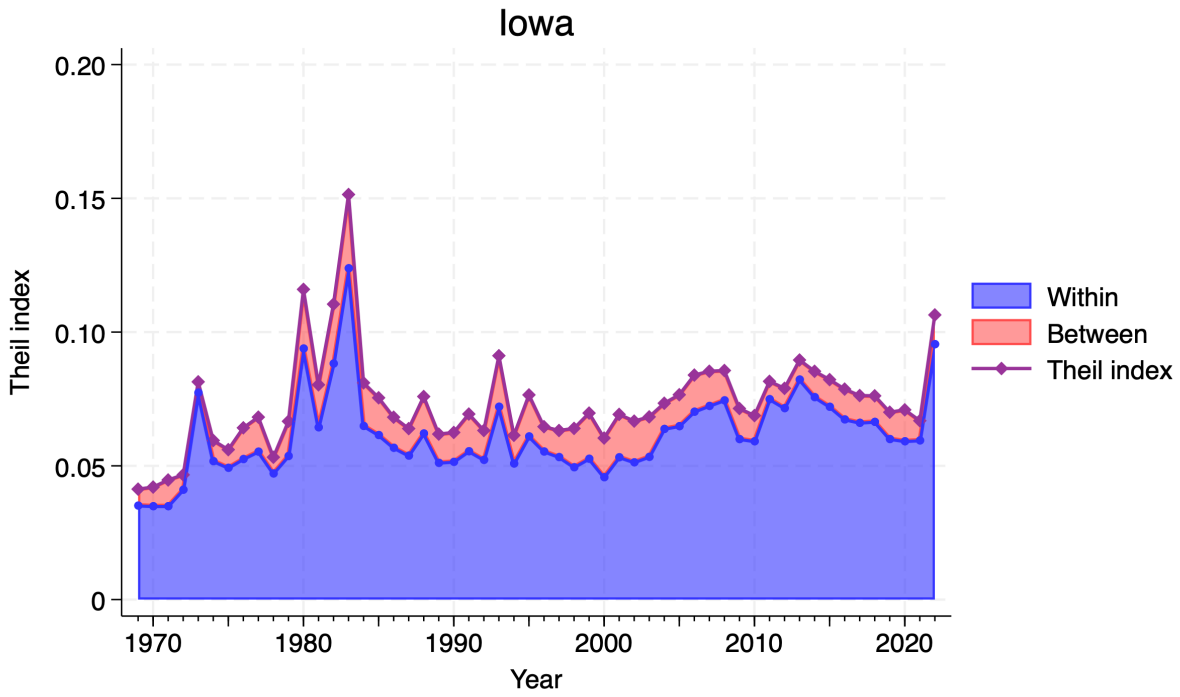

Georgia
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

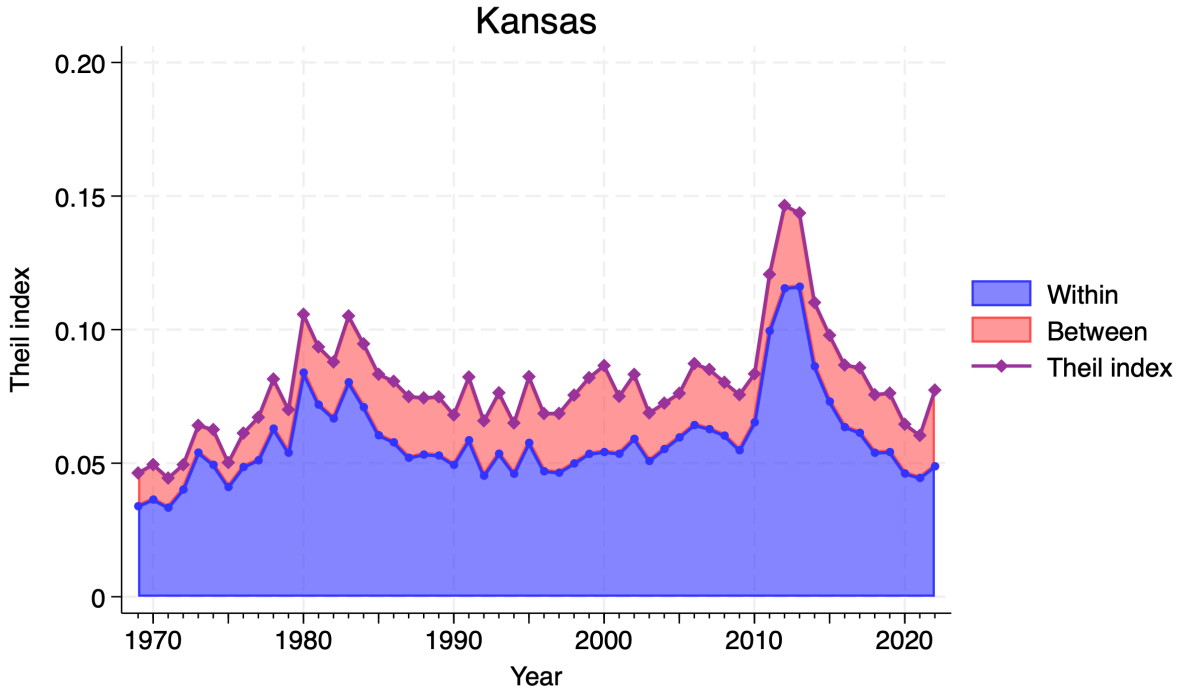

Hawaii
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

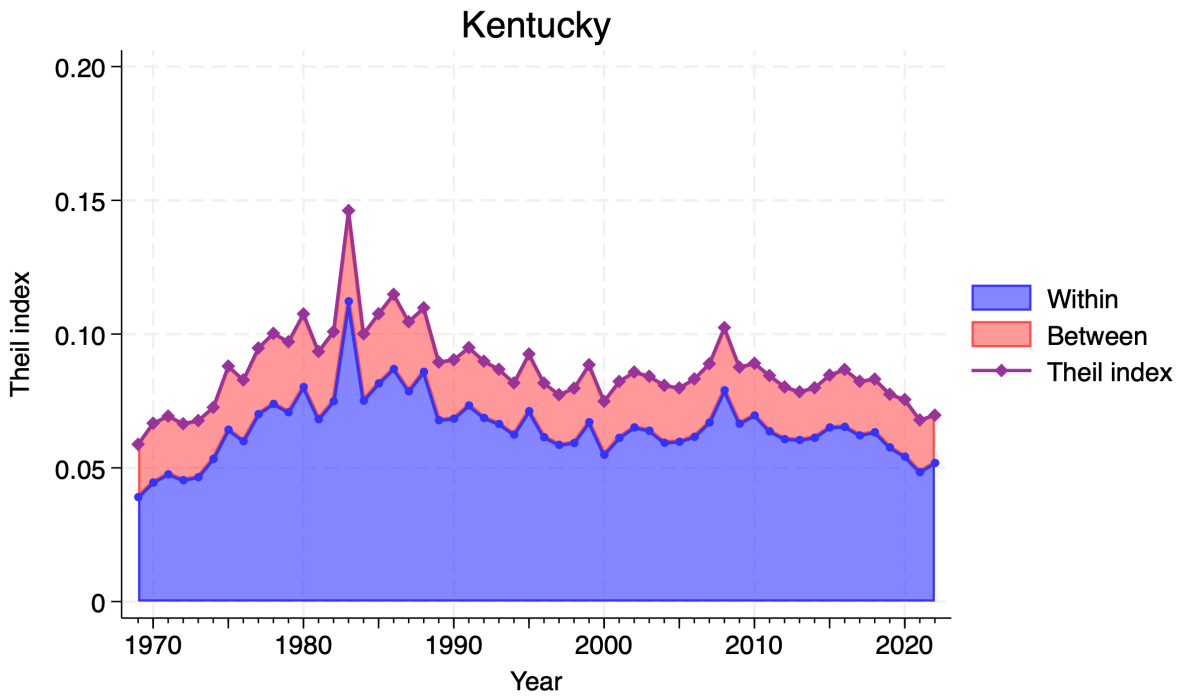

Idaho
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

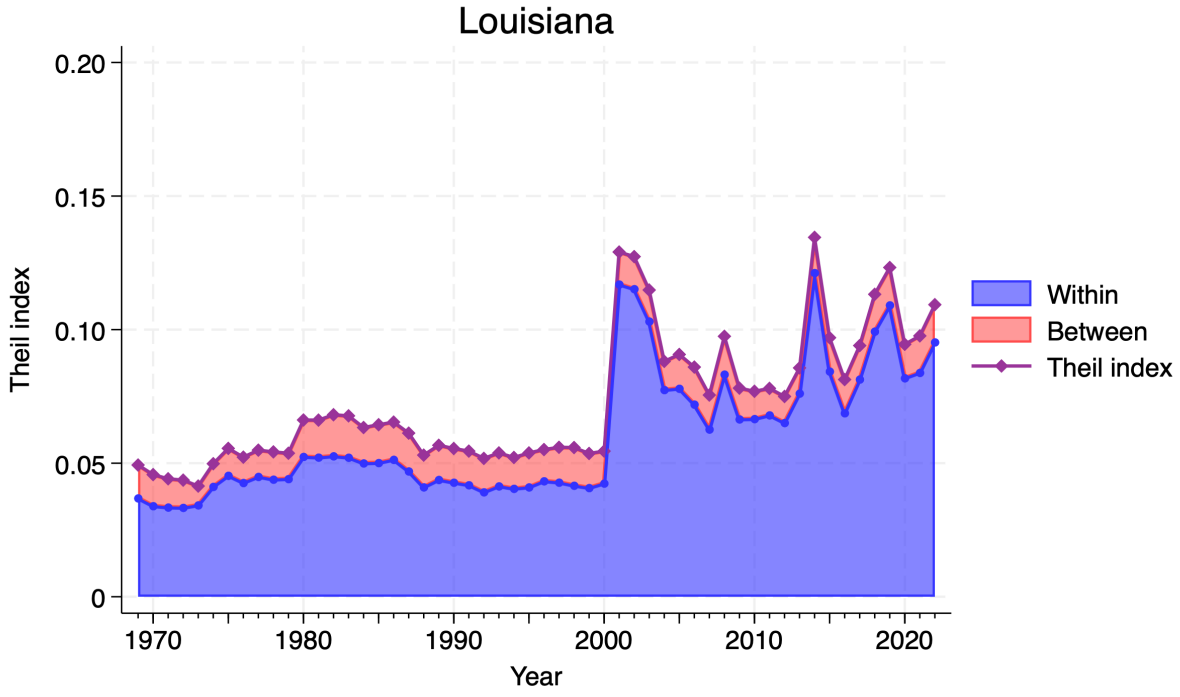

Illinois
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

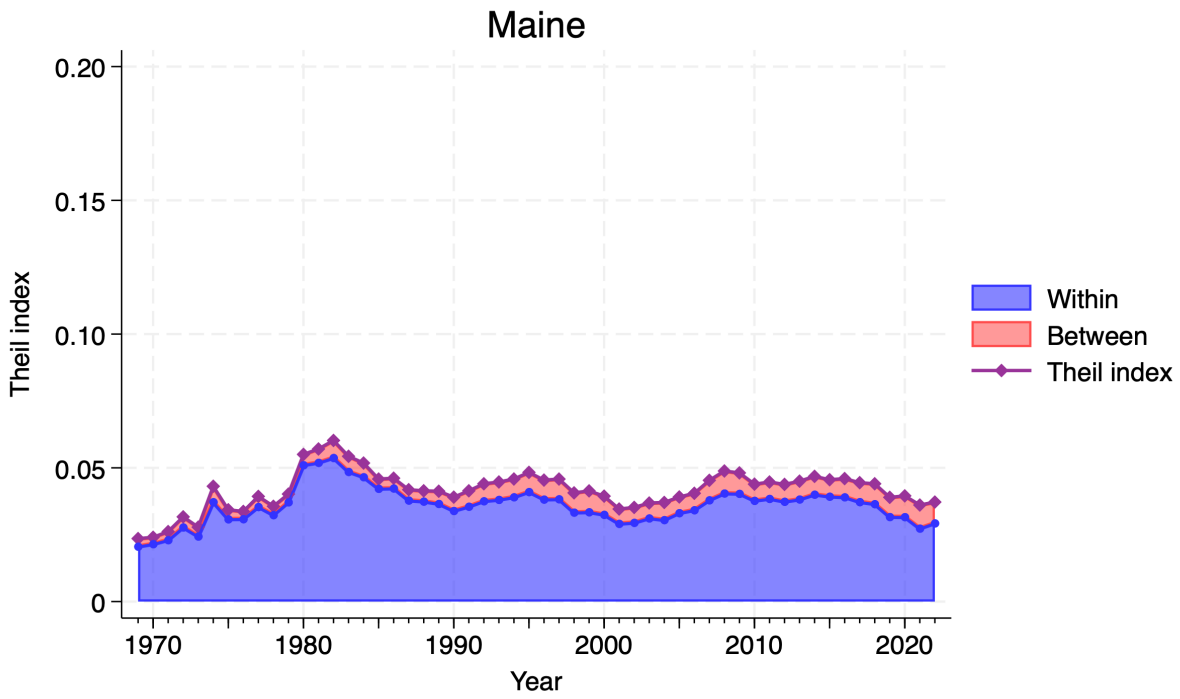

Indiana
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

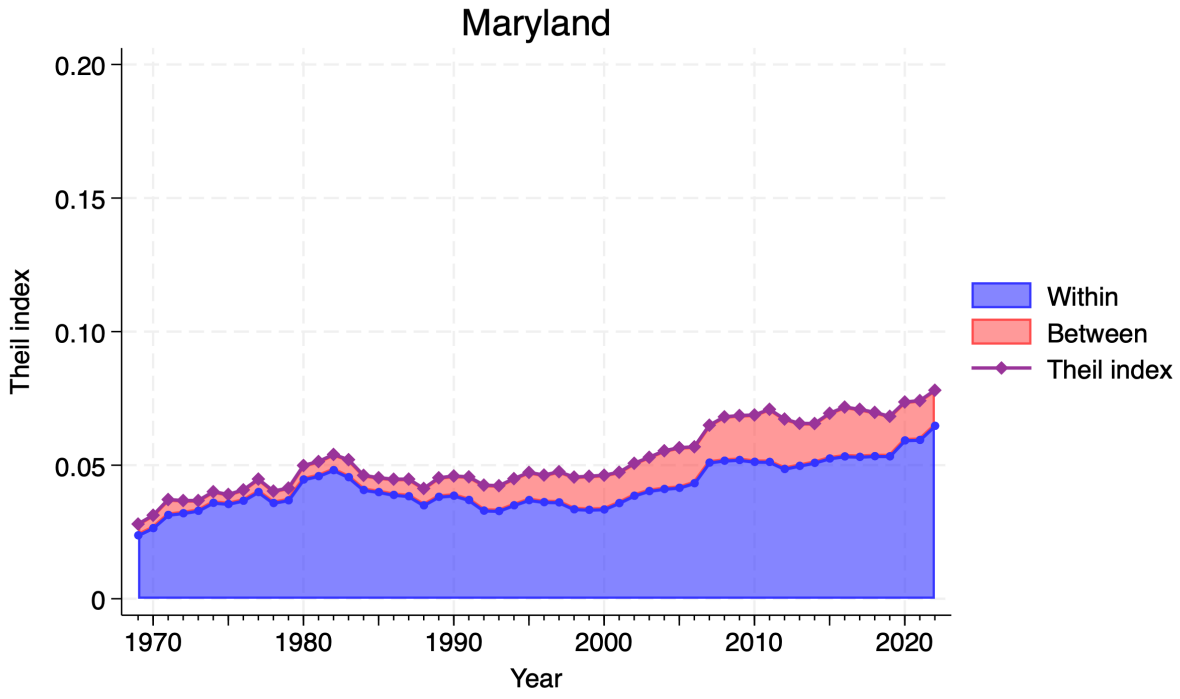

Iowa
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

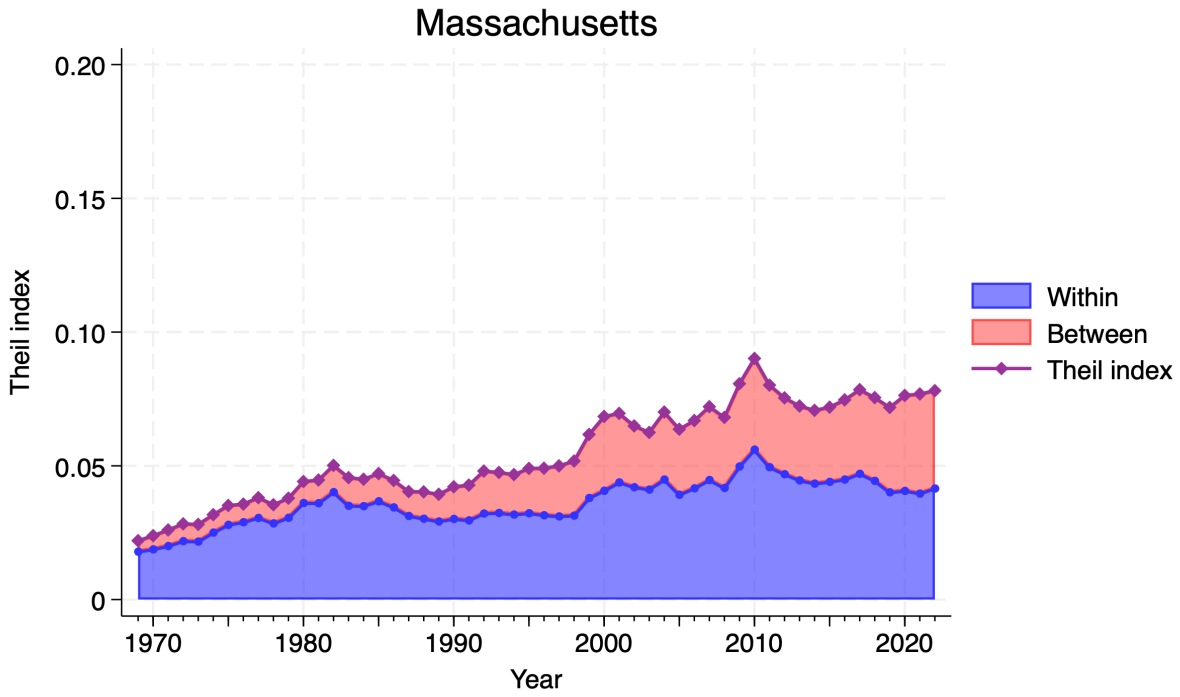

Kansas
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

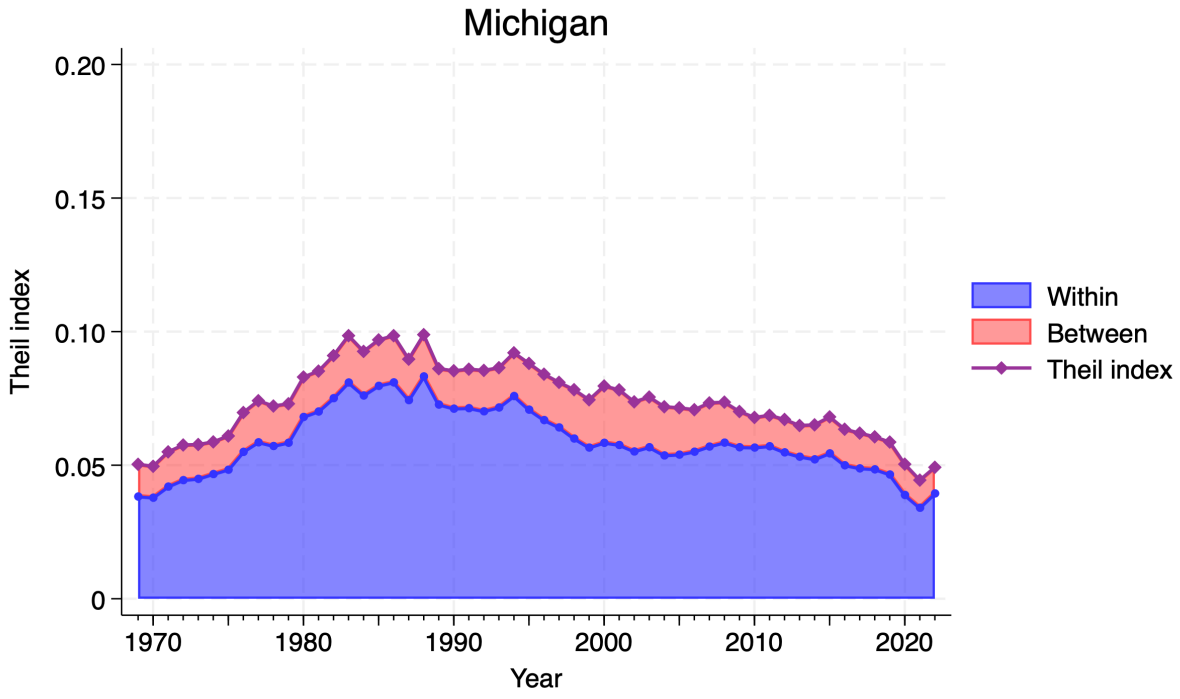

Kentucky
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

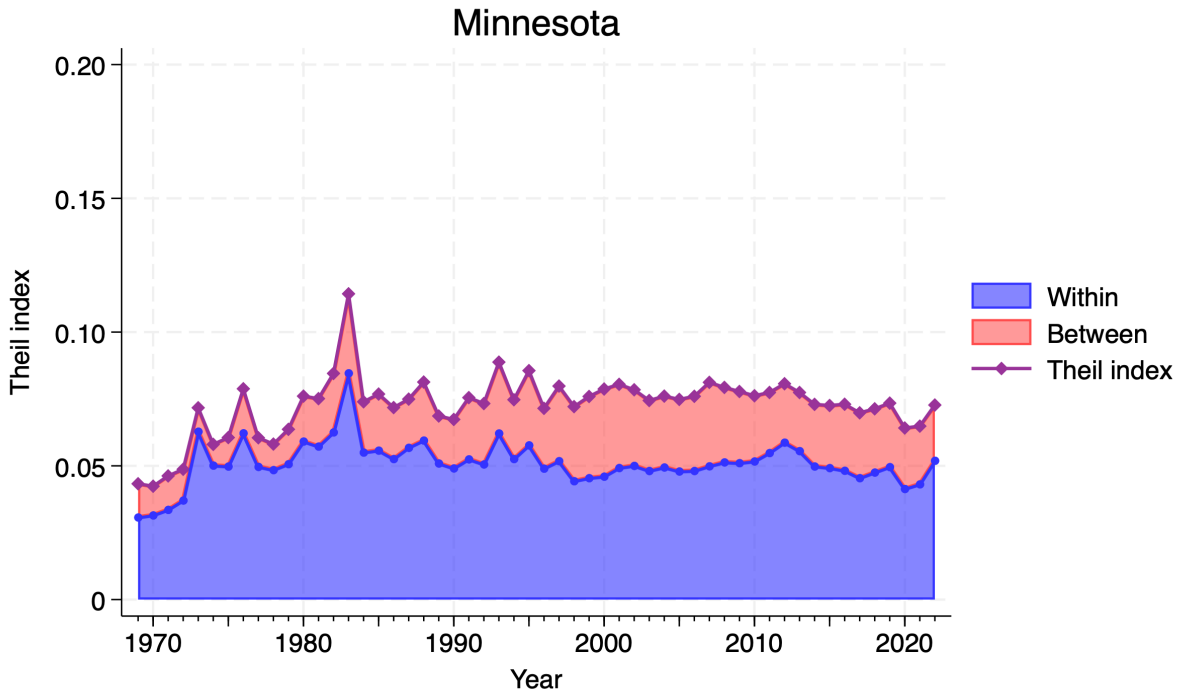

Louisiana
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

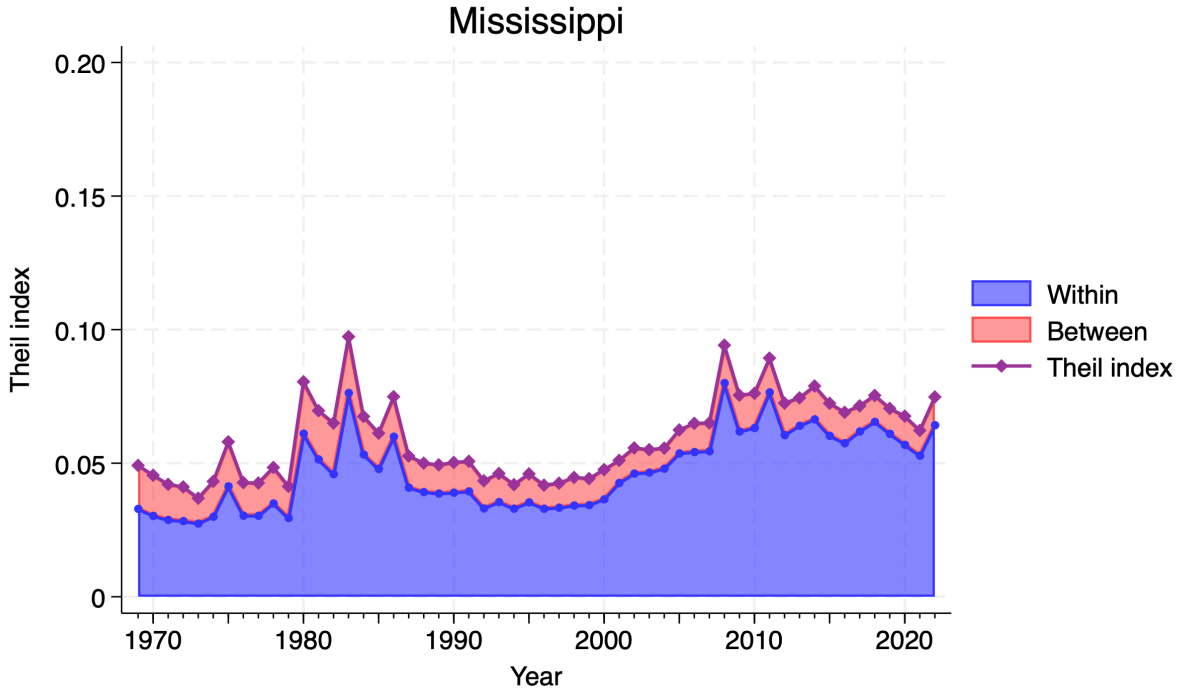

Maine
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

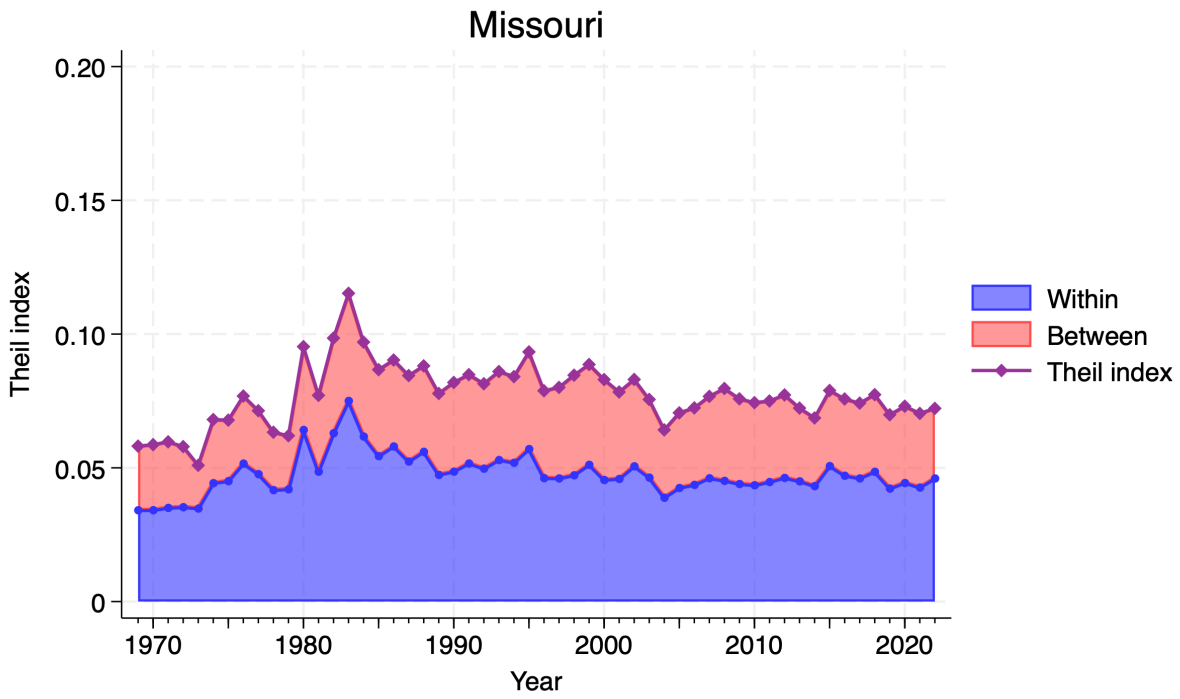

Maryland
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

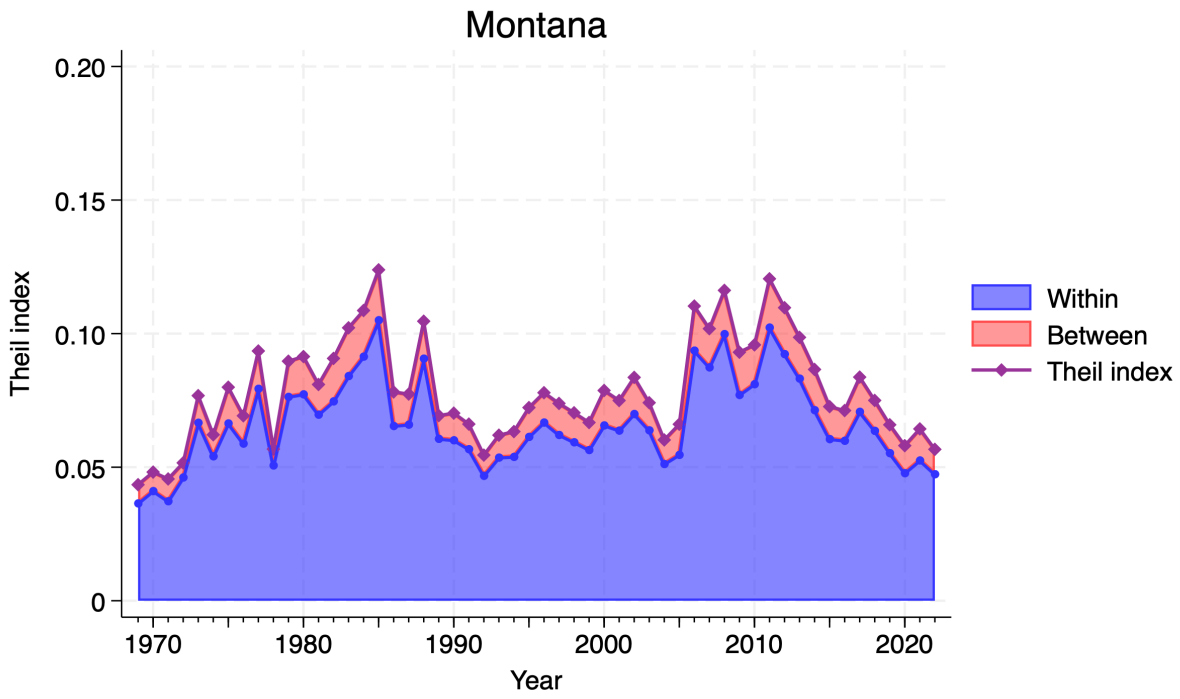

Massachusetts
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

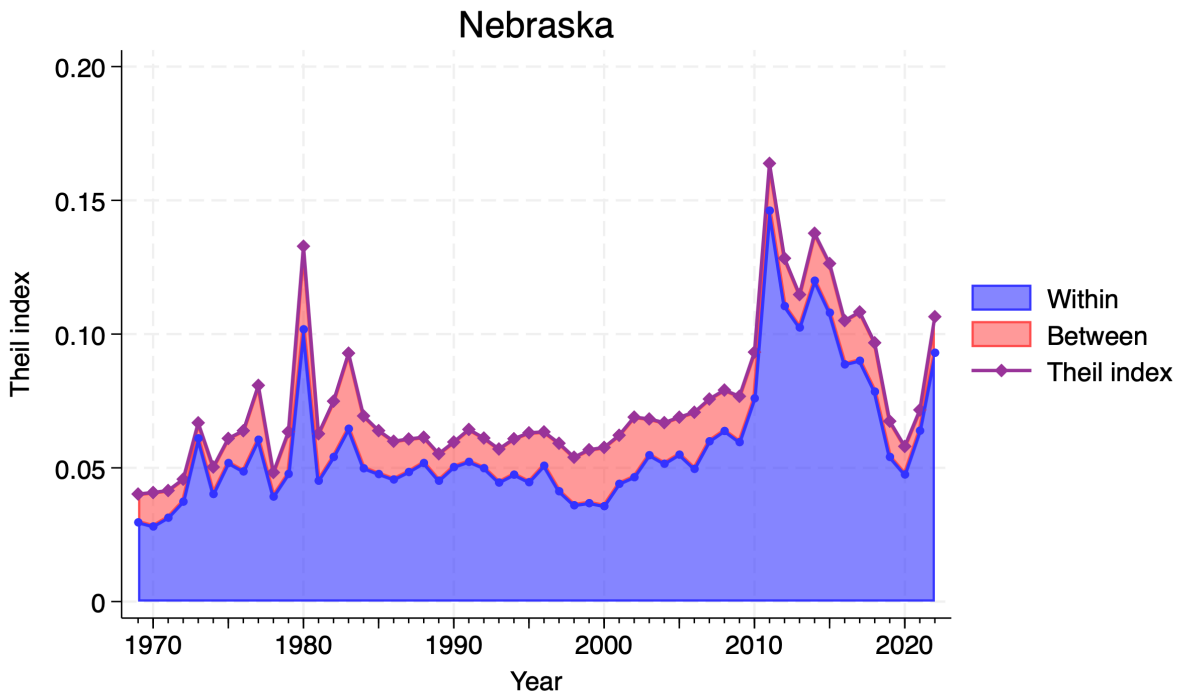

Michigan
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

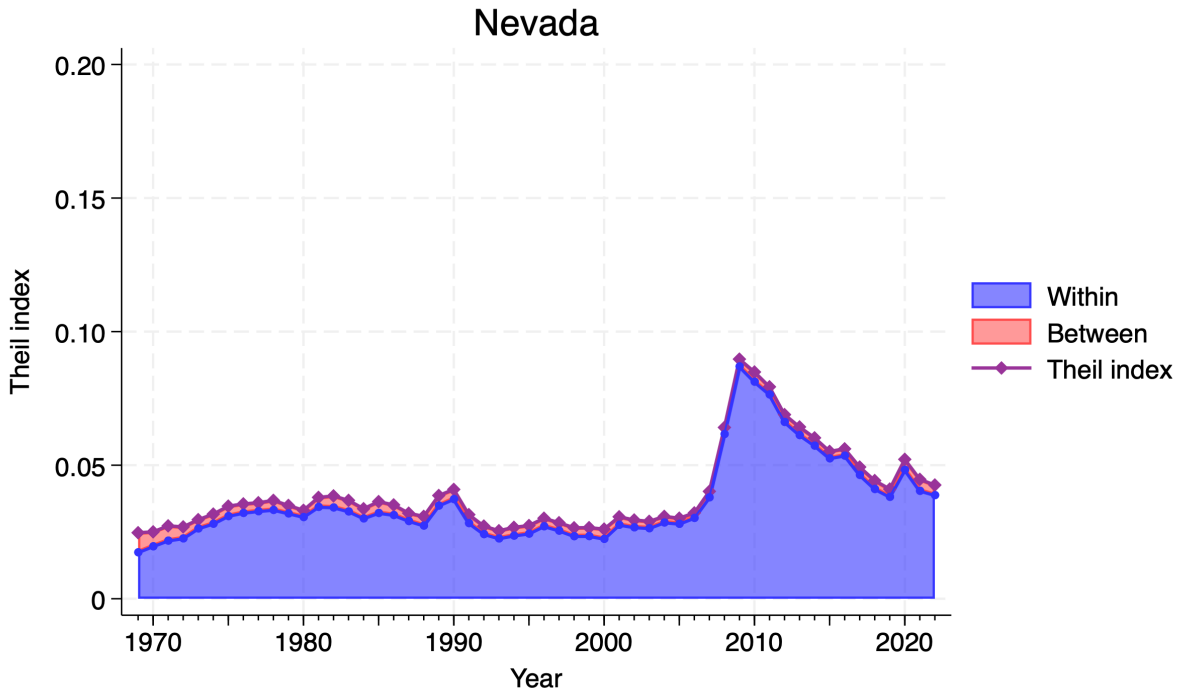

Minnesota
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

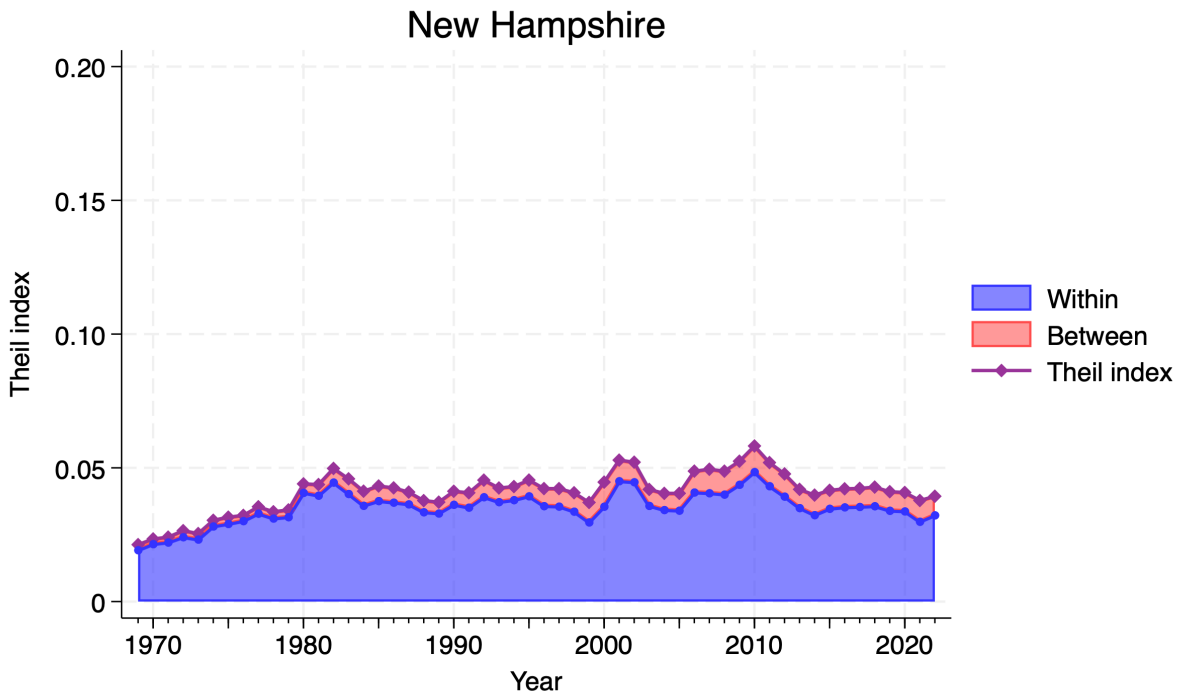

Mississippi
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

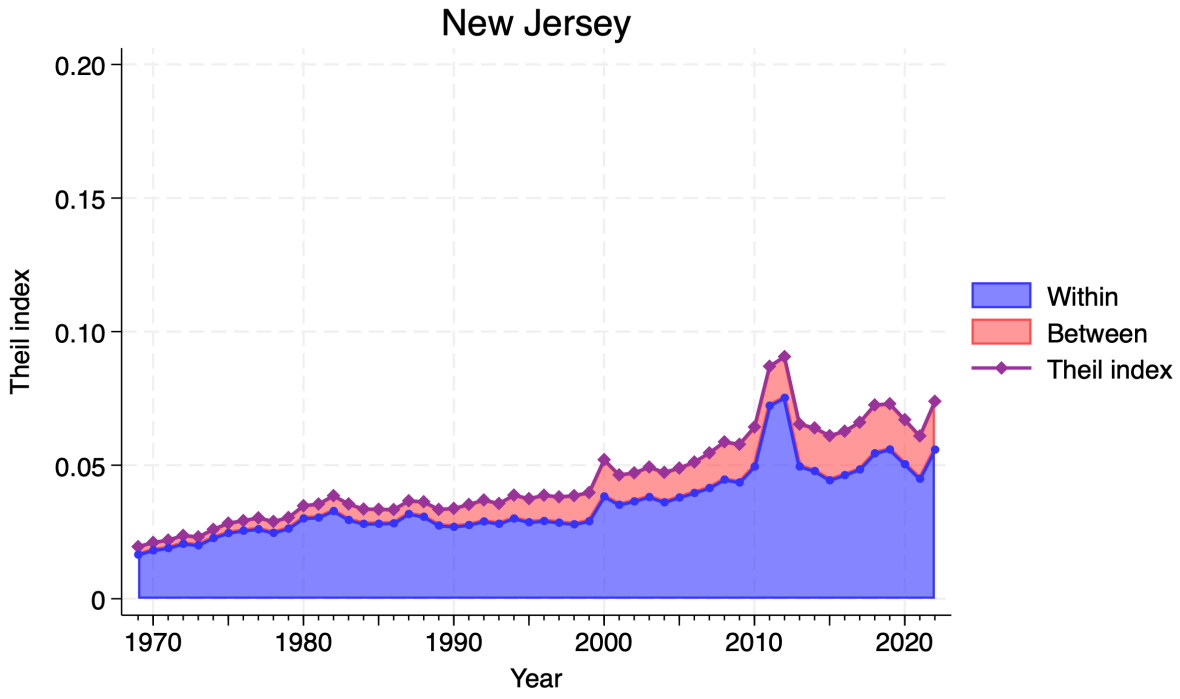

Missouri
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

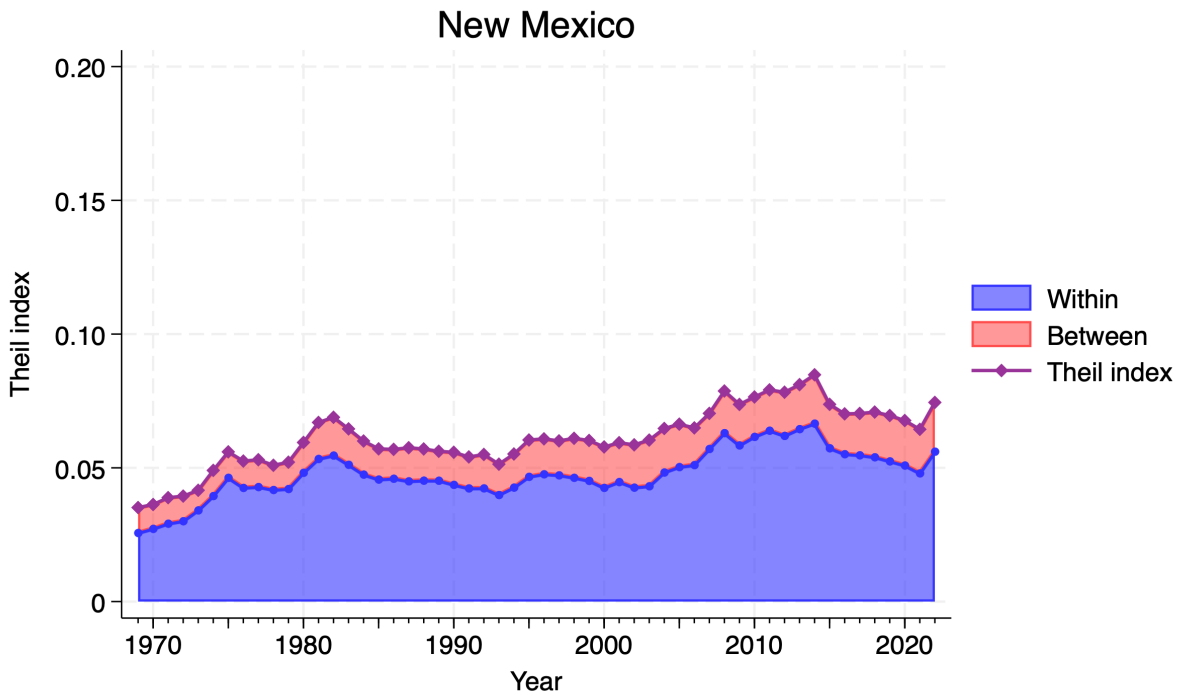

Montana
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

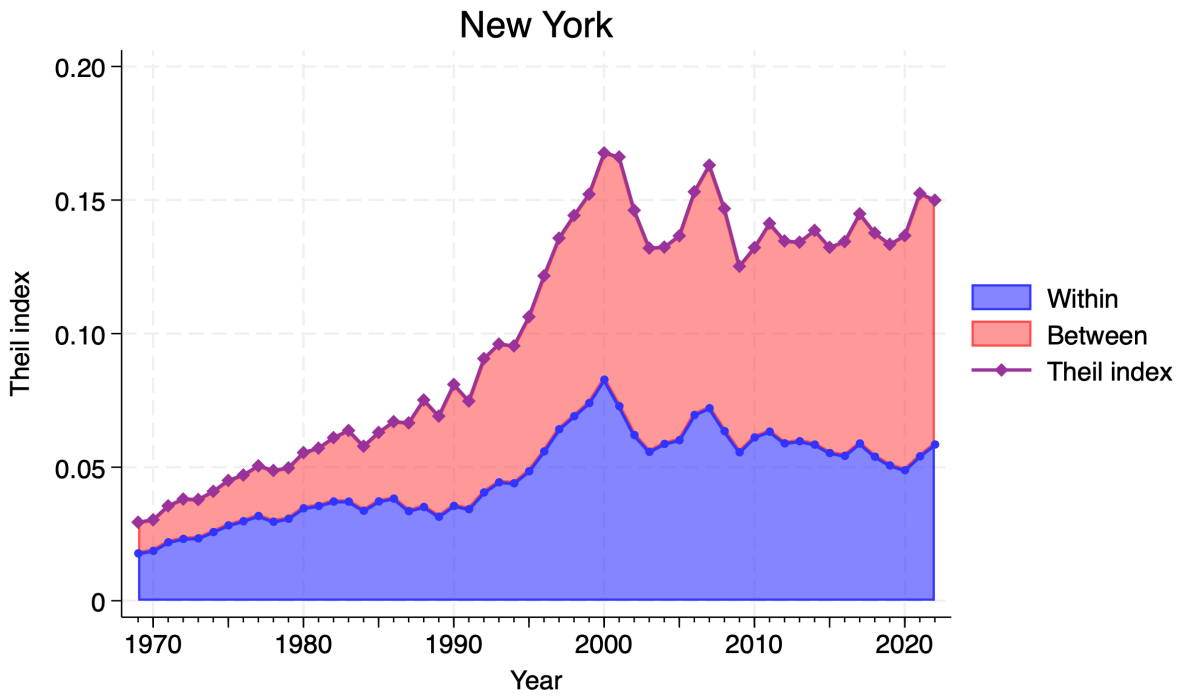

Nebraska
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

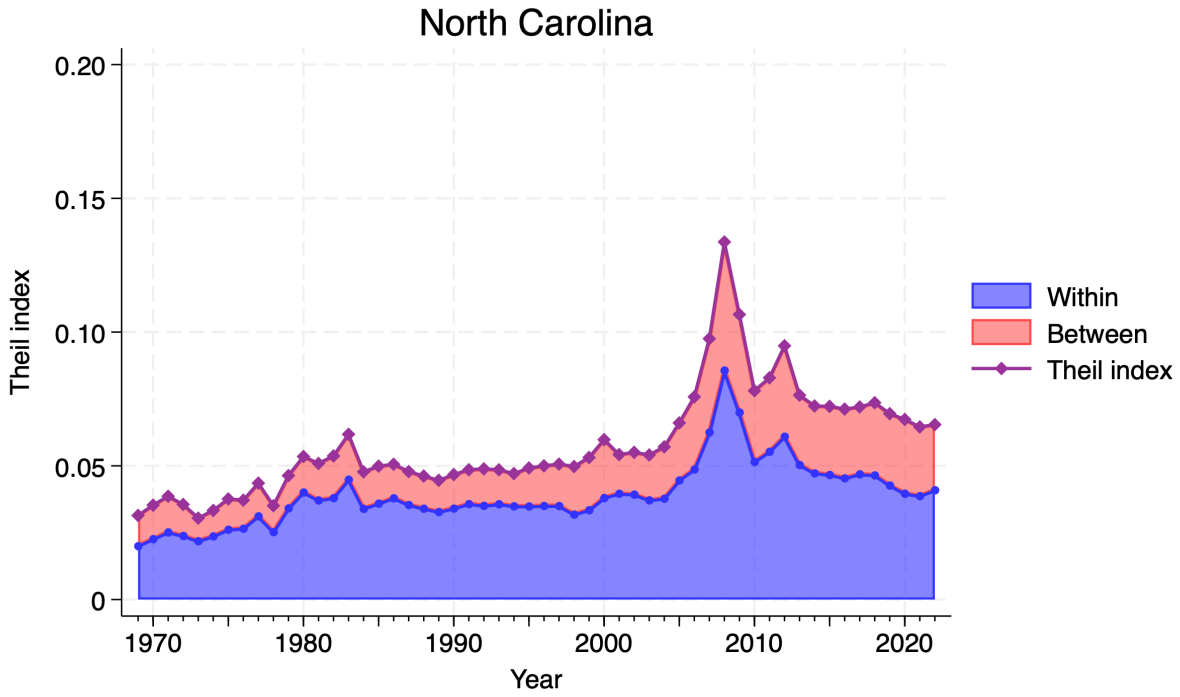

Nevada
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

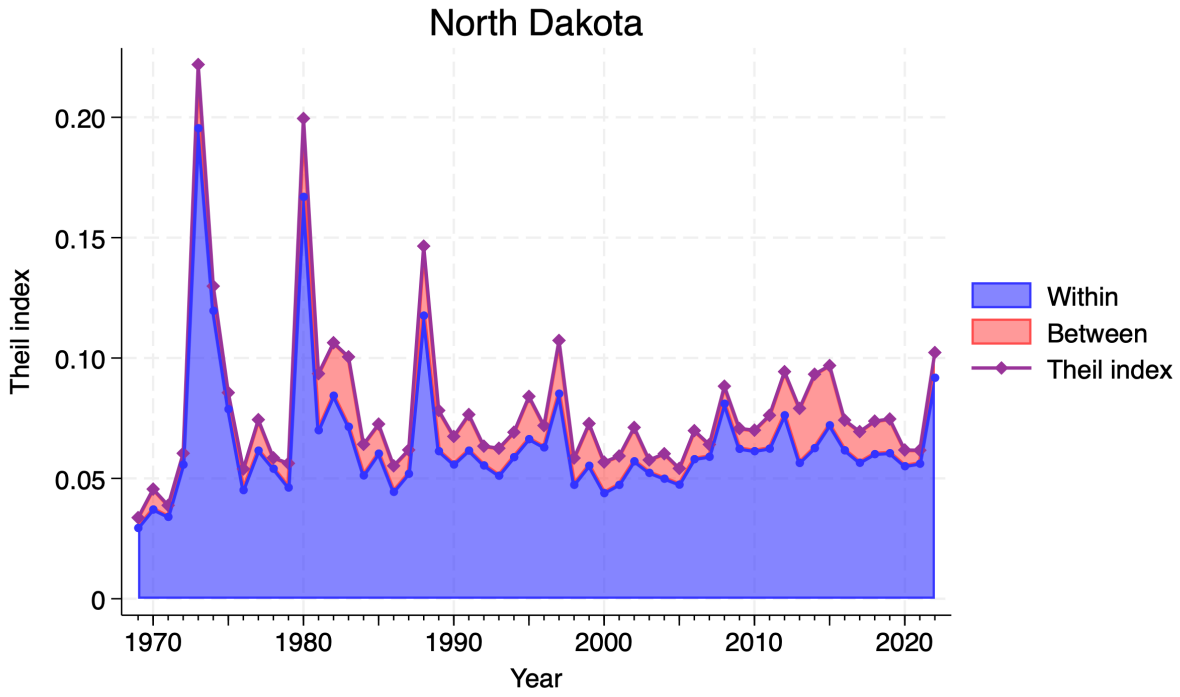

New Hampshire
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

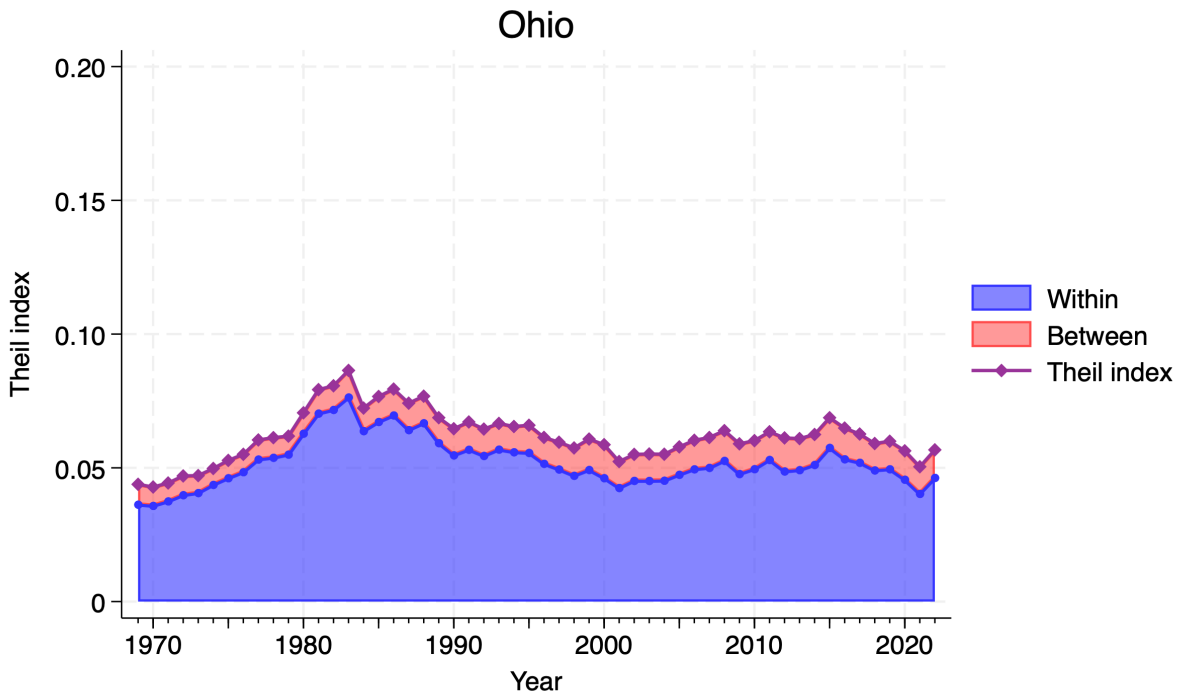

New Jersey
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

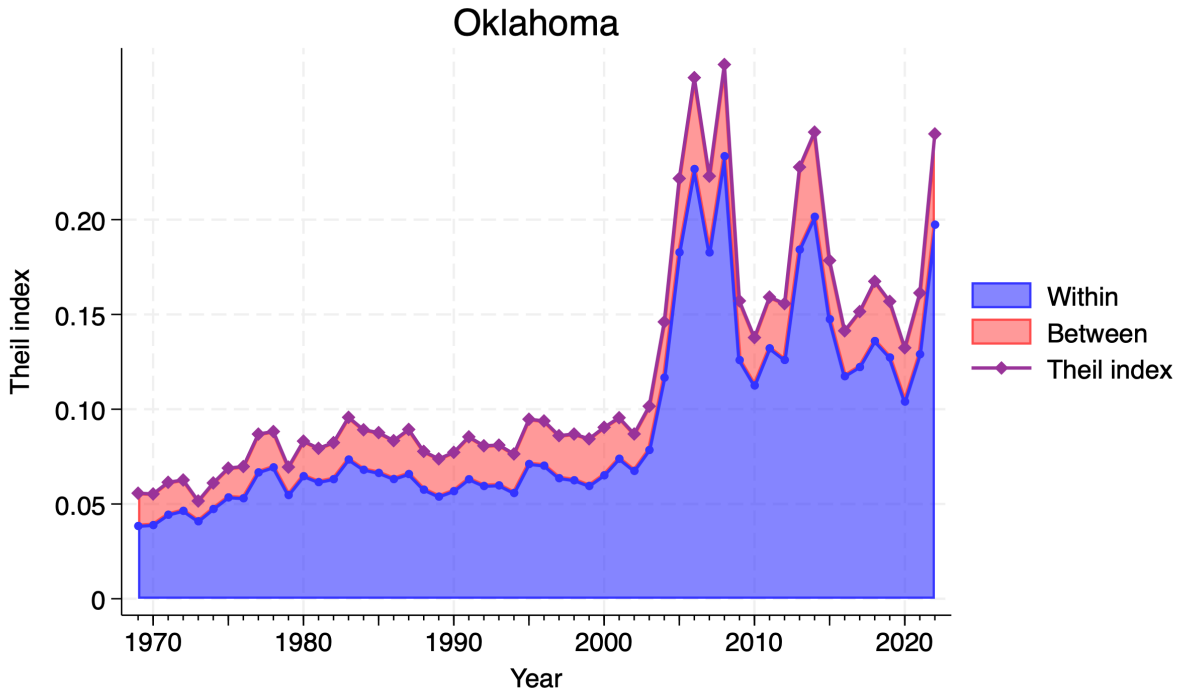

New Mexico
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

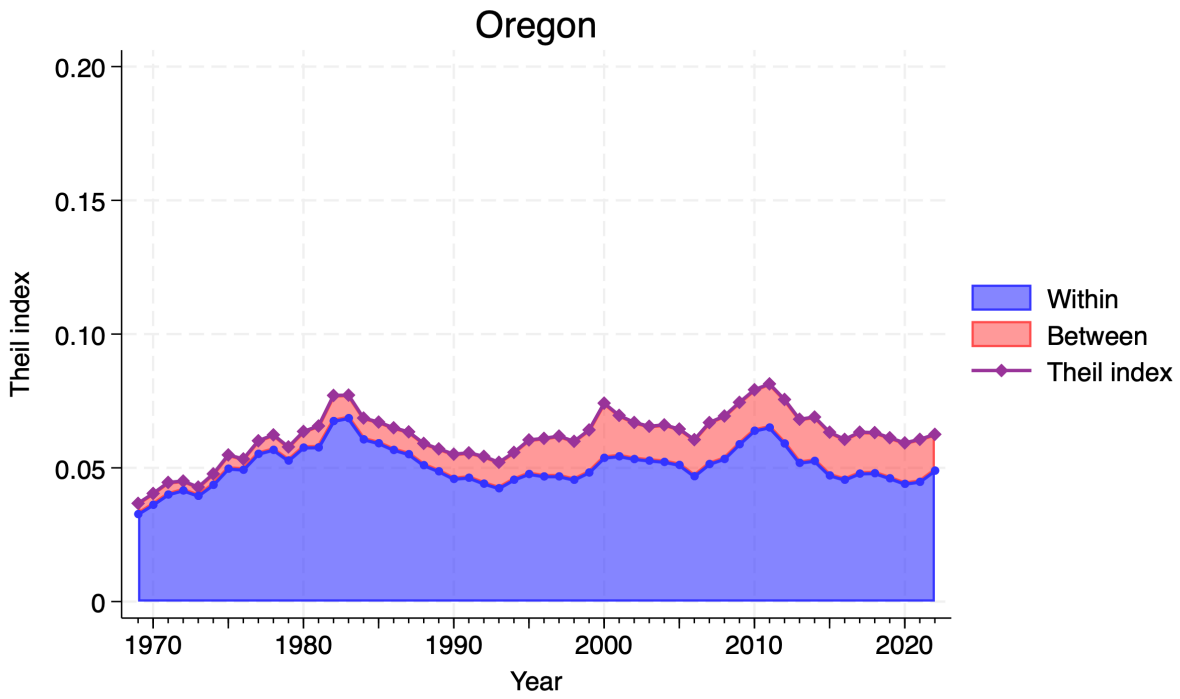

New York
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

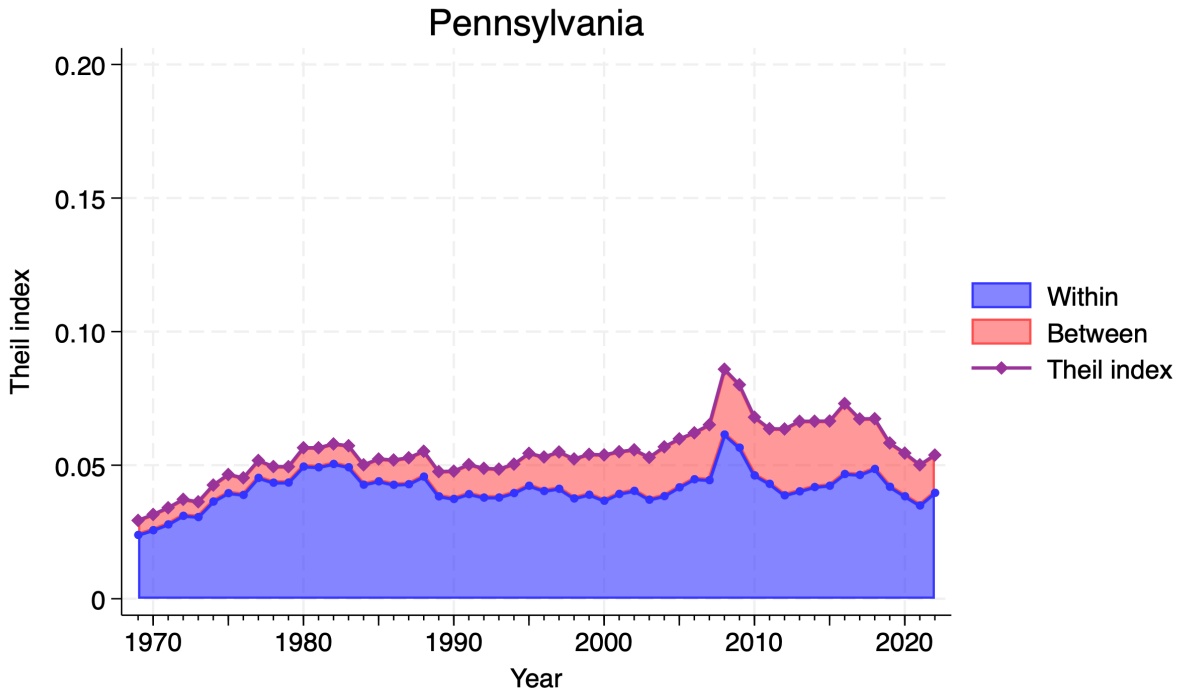

North Carolina
Theil index
0.20
0.15
0.10
0.05
0
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

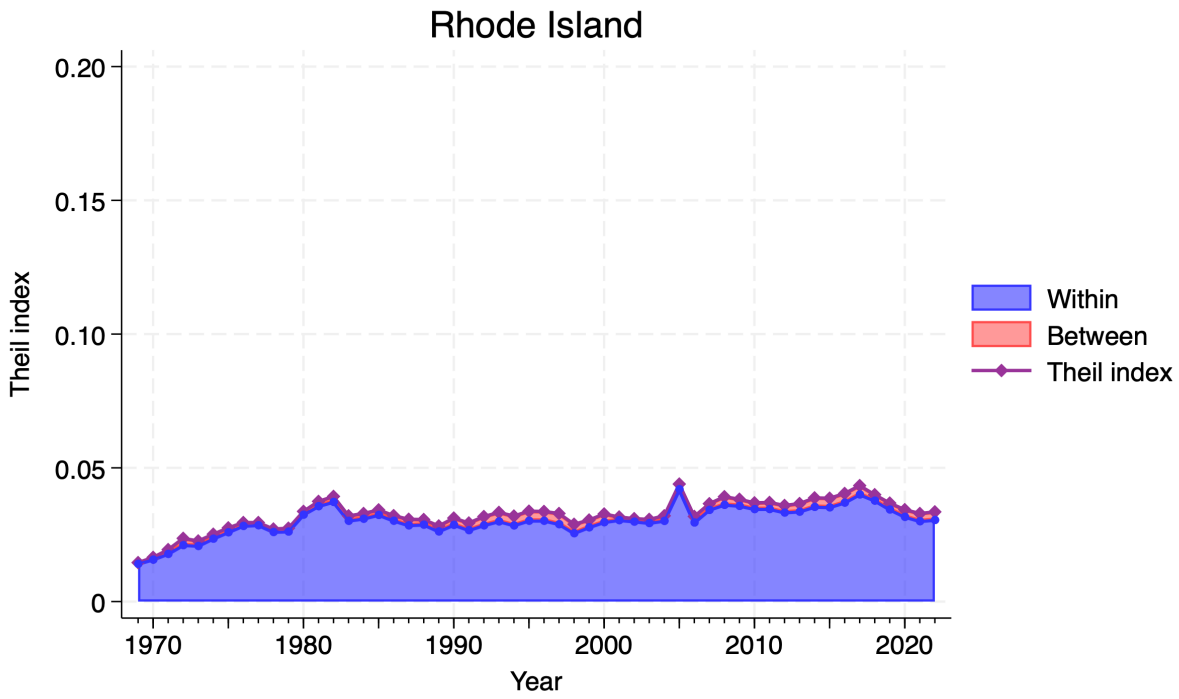

North Dakota
Theil index
0.20
0.15
0.10
0.05
0
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

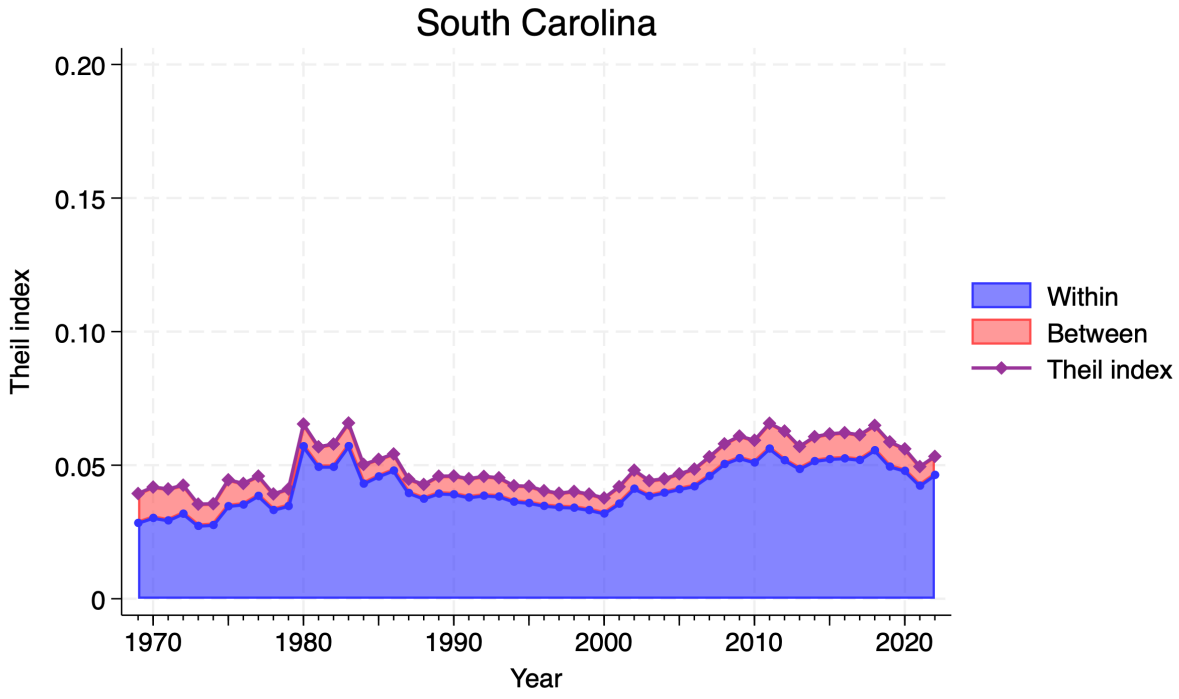

Ohio
Theil index
0.20
0.15
0.10
0.05
0
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

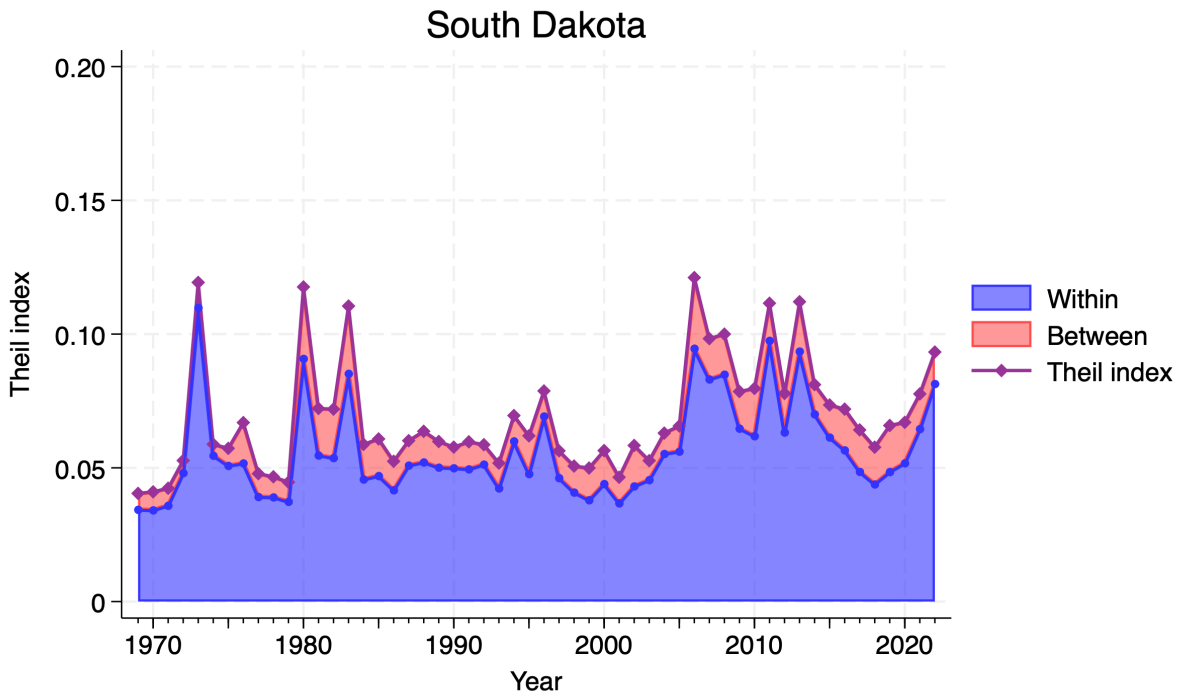

Oklahoma
Theil index
0.20
0.15
0.10
0.05
0
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

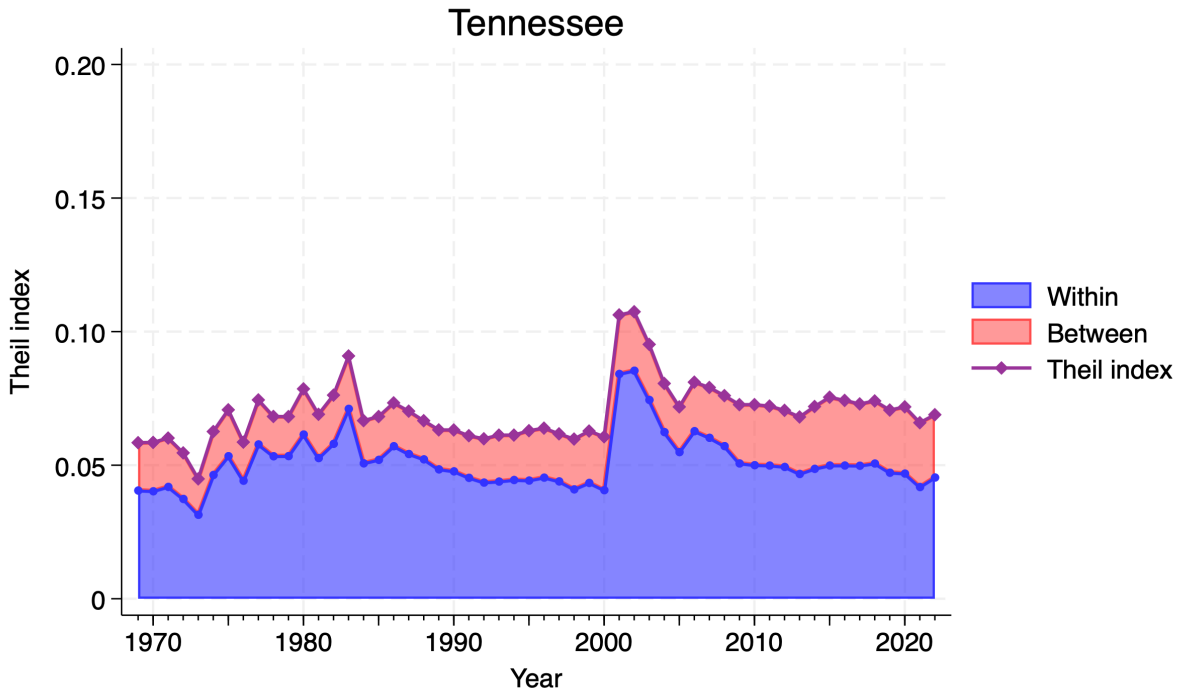

Oregon
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

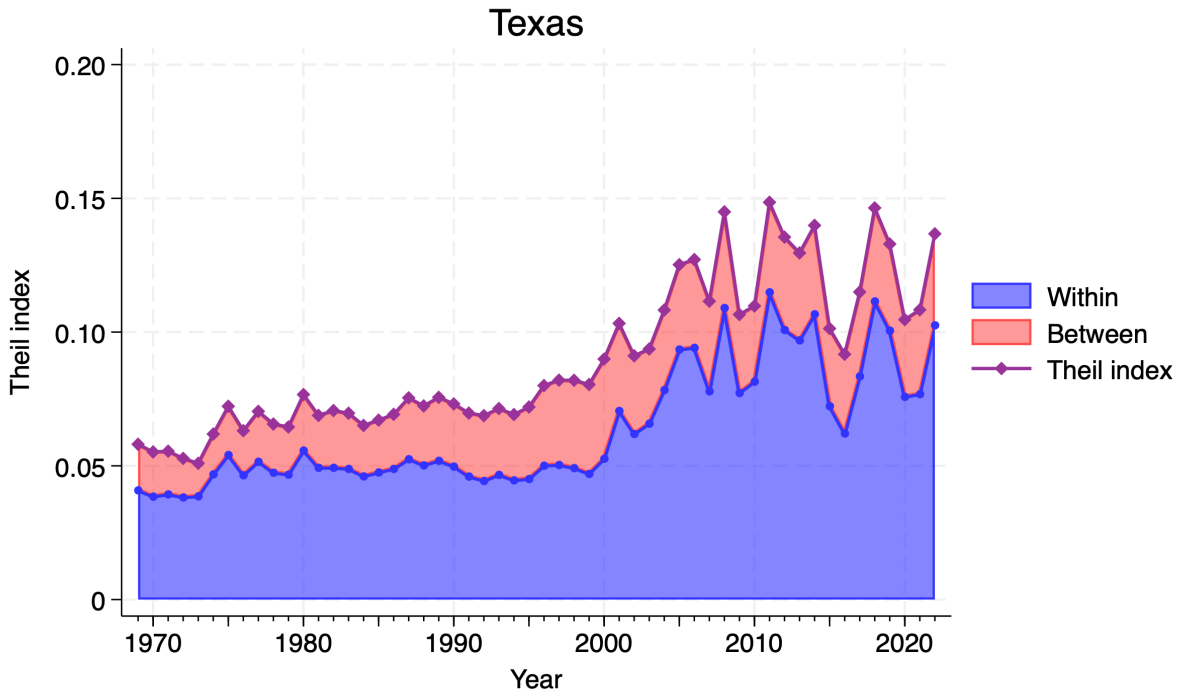

Pennsylvania
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

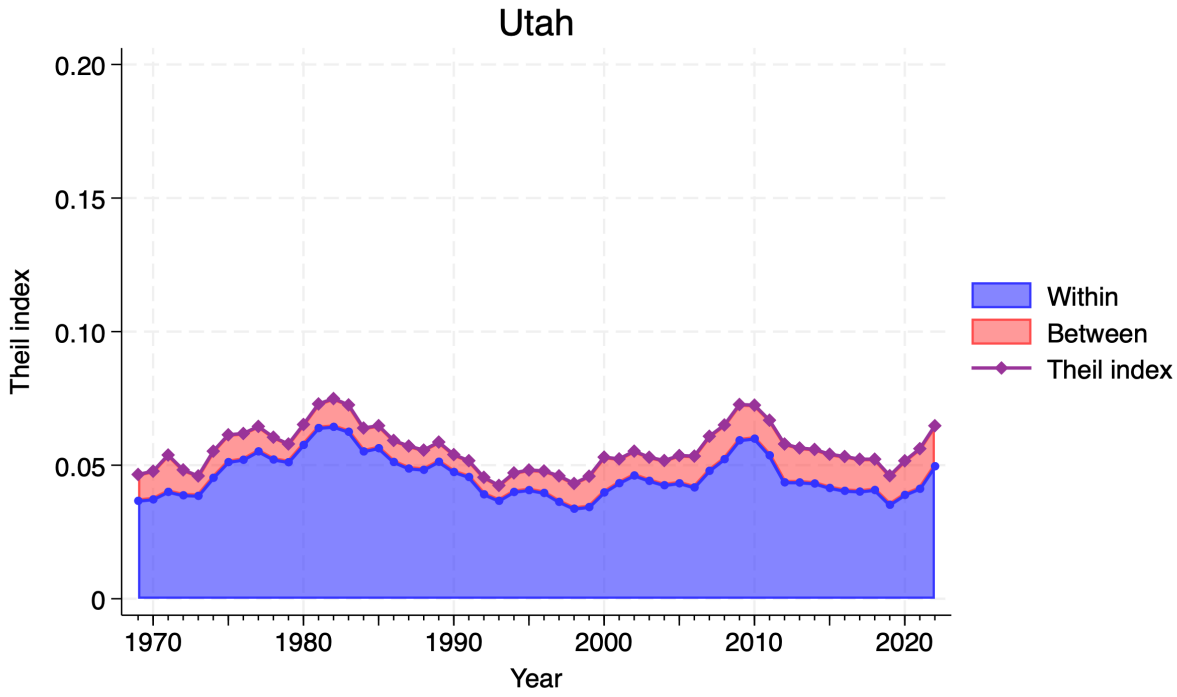

Rhode Island
Theil index
0.20
0.15
0.10
0.05
0
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

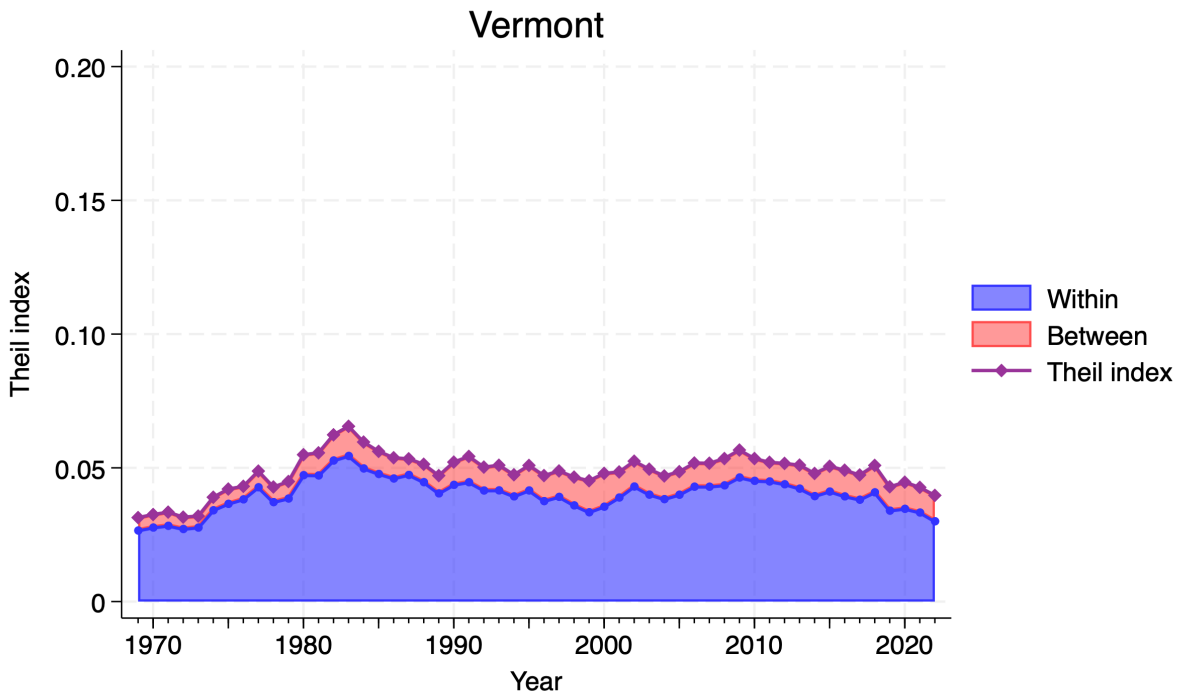

South Carolina
Theil index
0.20
0.15
0.10
0.05
0
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

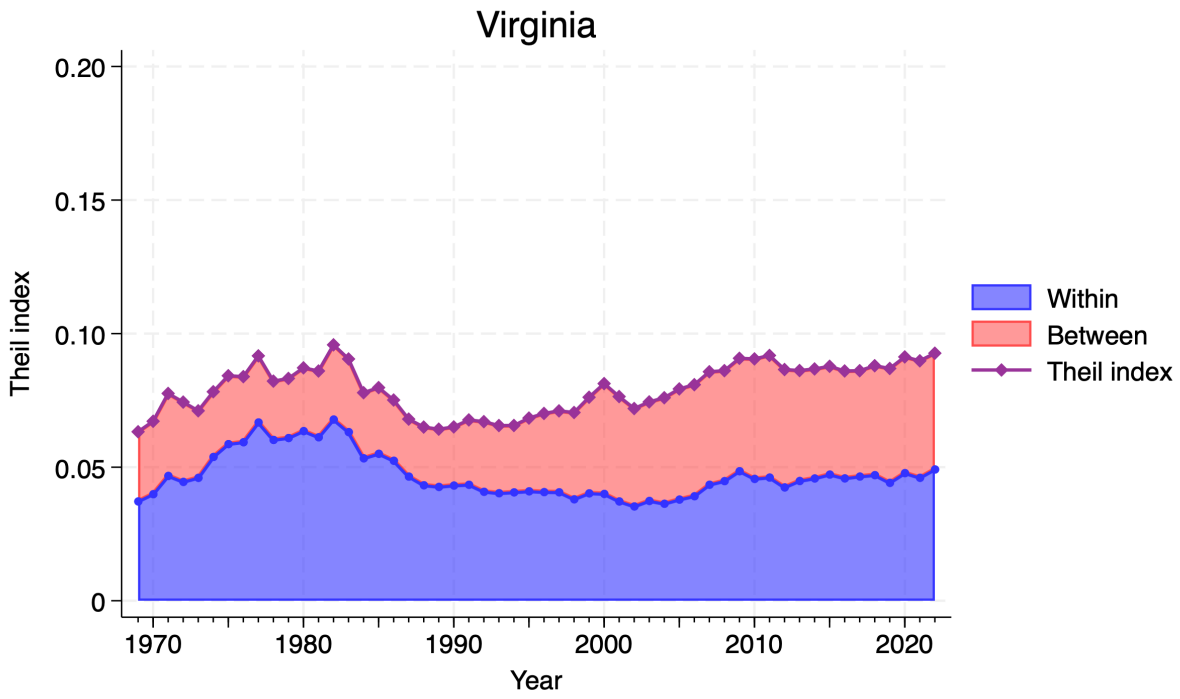

South Dakota
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

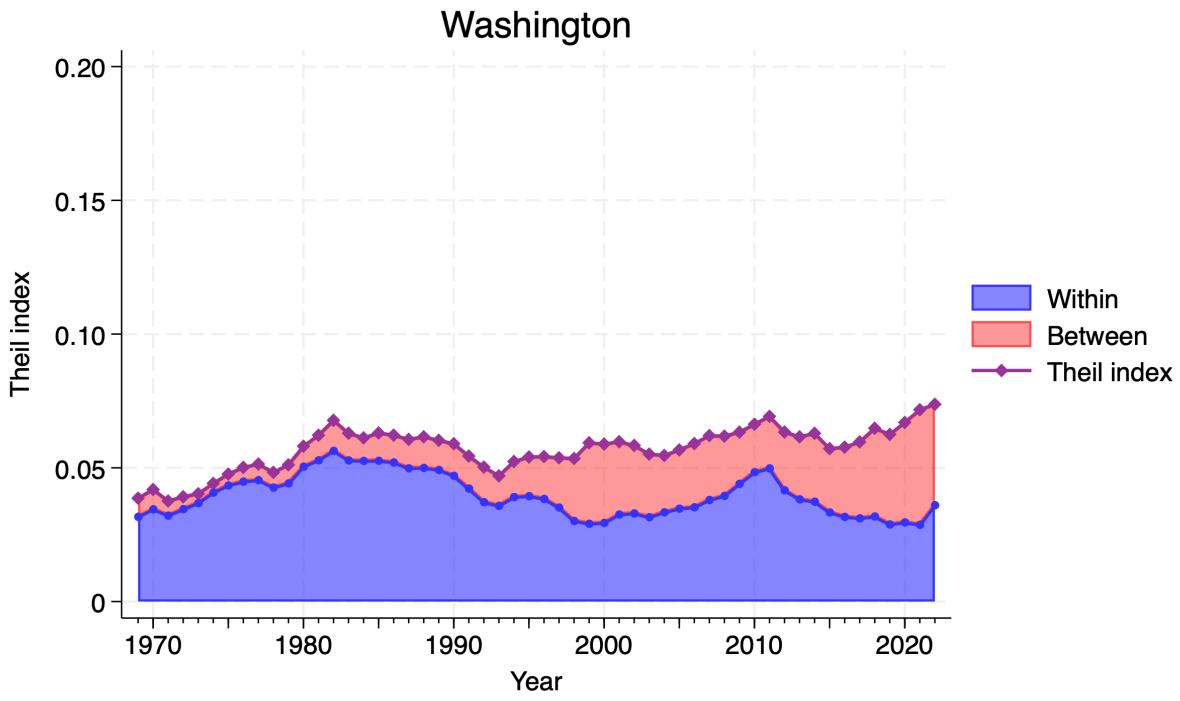

Tennessee
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

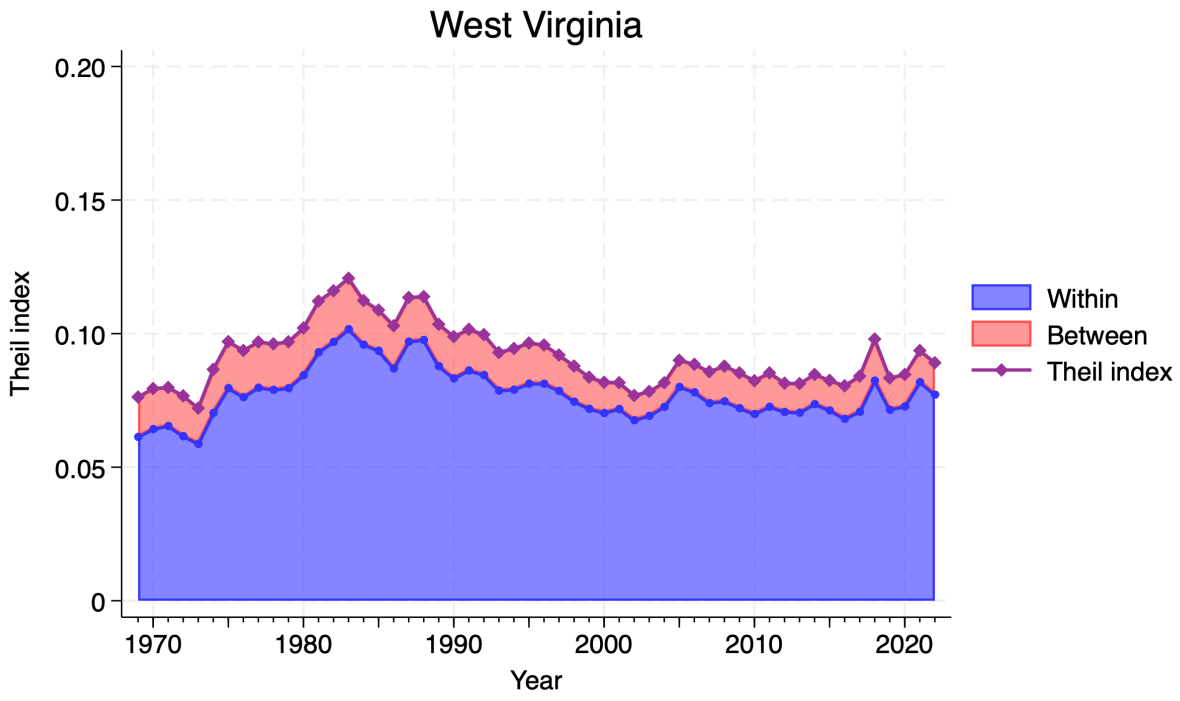

Texas
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

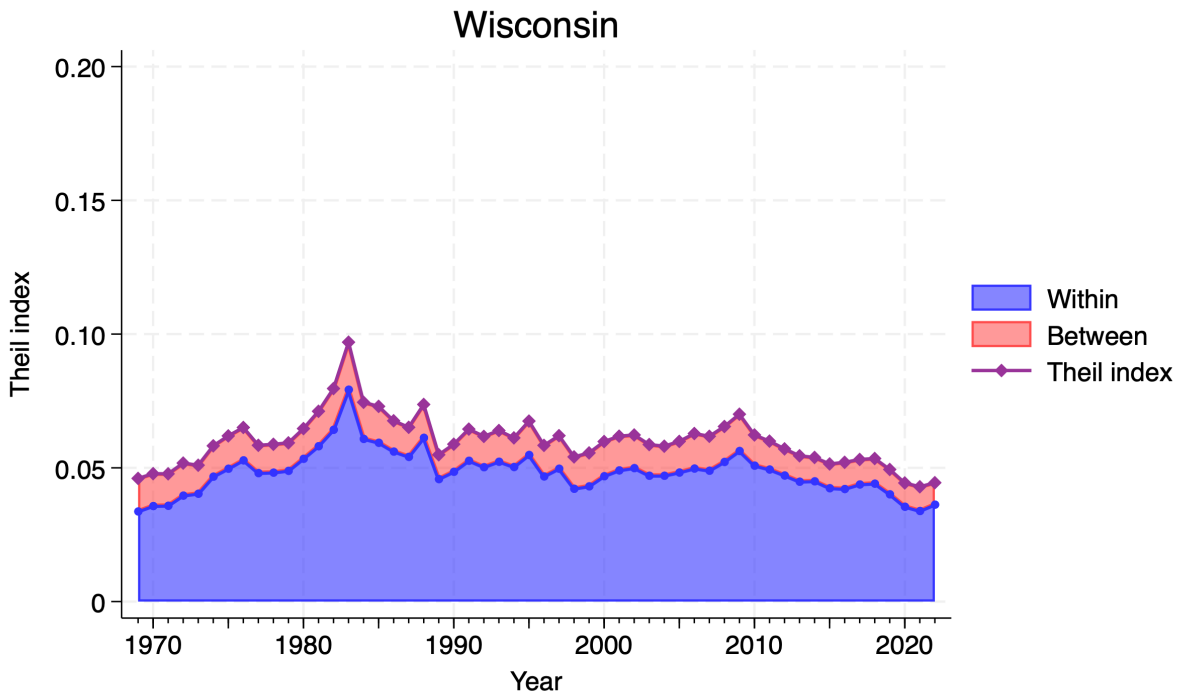

Utah
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

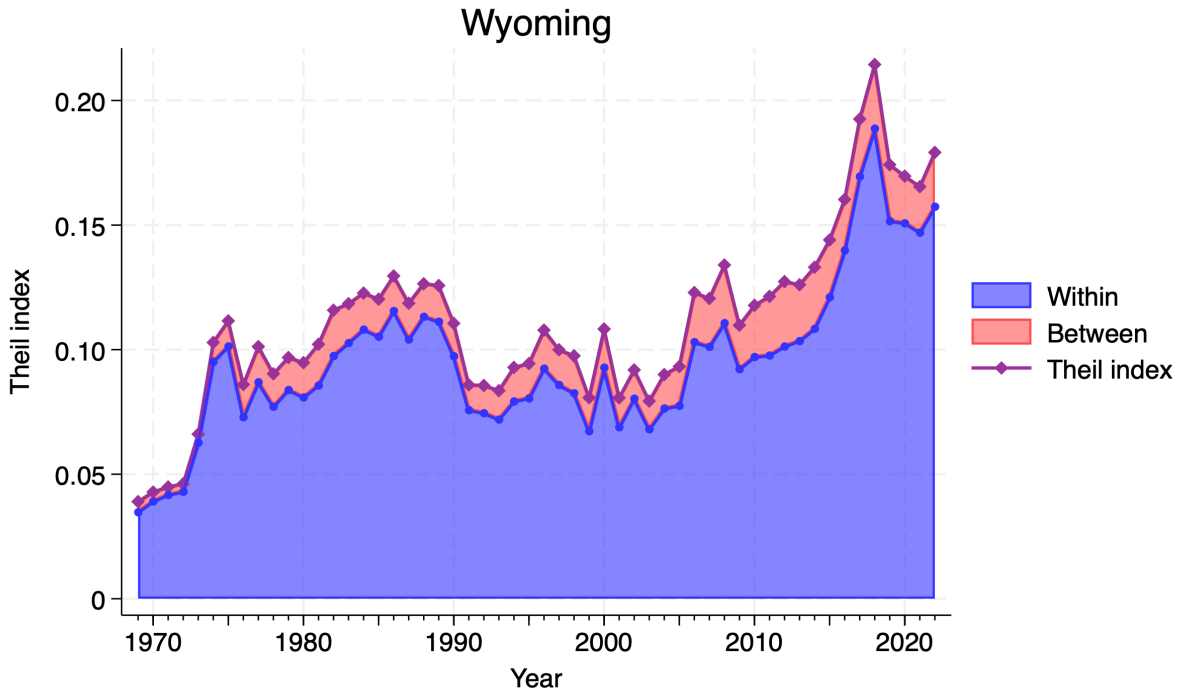

Vermont
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index


Virginia
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index


Washington
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index


West Virginia
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index


Wisconsin
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index


Wyoming
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

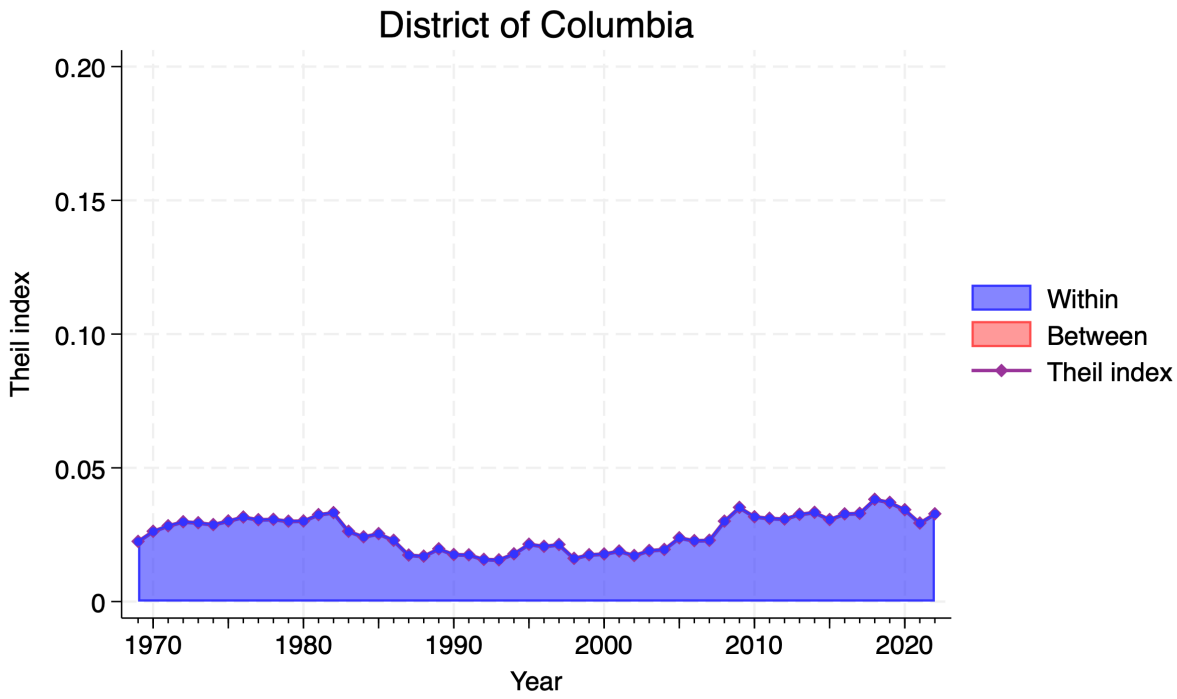

District of Columbia
0.20
0.15
0.10
0.05
0
Theil index
1970
1980
1990
2000
2010
2020
Year
Within
Between
Theil index

Table A1: Correlation of the state Theil index with the FSP top 1% share, by jurisdiction

| State | Levels | First diff. | State | Levels | First diff. |
|---|---|---|---|---|---|
| New York | 0.93 | 0.59 | Kansas | 0.36 | −0.05 |
| Massachusetts | 0.89 | 0.21 | Minnesota | 0.33 | −0.04 |
| Texas | 0.87 | 0.25 | Virginia | 0.32 | 0.00 |
| California | 0.84 | 0.71 | Nevada | 0.31 | −0.24 |
| Oklahoma | 0.81 | 0.05 | Nebraska | 0.30 | −0.10 |
| Georgia | 0.79 | −0.12 | Indiana | 0.30 | 0.02 |
| New Jersey | 0.78 | 0.04 | South Dakota | 0.27 | −0.11 |
| Maryland | 0.73 | −0.22 | Tennessee | 0.24 | −0.05 |
| New Mexico | 0.71 | 0.00 | Iowa | 0.23 | 0.15 |
| Arkansas | 0.66 | −0.12 | Hawaii | 0.20 | −0.12 |
| North Carolina | 0.63 | 0.00 | Vermont | 0.16 | −0.11 |
| Connecticut | 0.62 | −0.28 | Montana | 0.10 | 0.08 |
| Pennsylvania | 0.61 | −0.21 | Maine | 0.07 | −0.04 |
| Illinois | 0.60 | −0.31 | Missouri | 0.05 | 0.14 |
| Rhode Island | 0.59 | −0.05 | District of Columbia | 0.00 | −0.22 |
| Colorado | 0.59 | −0.12 | Ohio | −0.02 | −0.05 |
| Louisiana | 0.57 | 0.19 | Arizona | −0.03 | −0.23 |
| Wyoming | 0.55 | 0.10 | North Dakota | −0.07 | 0.15 |
| Washington | 0.54 | 0.10 | Kentucky | −0.12 | 0.23 |
| Florida | 0.47 | −0.07 | Alaska | −0.13 | −0.22 |
| New Hampshire | 0.47 | −0.11 | Utah | −0.16 | −0.19 |
| Idaho | 0.46 | −0.07 | West Virginia | −0.21 | 0.03 |
| Mississippi | 0.45 | 0.07 | Wisconsin | −0.29 | 0.07 |
| Oregon | 0.42 | 0.03 | Michigan | −0.30 | −0.04 |
| Alabama | 0.39 | −0.05 | Delaware | −0.39 | −0.35 |
| South Carolina | 0.36 | −0.12 | | | |

**Note:** Pearson correlations between the total Theil index and the FSP top 1% share within each jurisdiction, 1969–2022 ($n = 54$). First differences are annual changes. Medians: 0.36 in levels and −0.05 in first differences. FSP = Frank et al. (2015).