

# How To Implement and Interpret Name-Based Estimators of Intergenerational Mobility

Andrea Del Pizzo

8 January 2026

## Motivation: Name-Based Estimators

- ▶ A growing literature uses names or surnames as instruments in TSLS frameworks to assess intergenerational persistence due to *data constraints*. (Olivetti and Paserman, 2015)
  - the absence of direct family links across generations (Collado et al., 2012; Barone and Mocetti, 2021; Belloc et al., 2024)
  - concerns about measurement error in observed parental income (Clark, 2014)
- ▶ *Empirically*, surname-based estimators often deliver **very high** persistence estimates:
  - *alternative measures* of the standard IGE?
  - *different estimand*?
- ▶ *However*, surname-based estimates are also very sensitive to various sources of **measurement error**
  - Attenuation Bias

# Surname-based estimators: empirical objects

- ▶ **Standard benchmark: parent-child persistence**

- Intergenerational elasticity / correlation:

$$y_{it} = \alpha + \beta^{PC} y_{it-1} + \varepsilon_{it}$$

- ▶ **Surname-based persistence**

- Surname regression:

$$y_{it} = \alpha + \beta^S \bar{y}_{s(i)t-1} + u_{it}$$

where  $\bar{y}_{s(i)t-1}$  is the surname average in the parent generation

- Equivalent formulation → use surname as IV for father income

- ▶ The surname average is the fitted value from the first stage

# Outline

## ① Identification in Perfect Data

- Identification
- Name Frequency
- Weighting Structure

## ② Consequences of Imperfect Data

- Imperfect Linking
- Sampling the Census
- Lack of Overlap

## ③ Regional Analysis and Conditioning on Covariates

- Conditioning on Geography
- Comparison Across Areas

## ④ Conclusion

# Data: the “perfect” benchmark

- ▶ We use restricted-access IPUMS-US Census data (1920–1940) with individual names.
  - Sample: males aged 30–40 in 1940 who are successfully linked to their fathers.
  - Child outcomes are observed in 1940; father outcomes are observed in 1920.
- ▶ To minimize *measurement error*, we impose three benchmark assumptions, defining a **“perfect” data environment**:
  1. **Perfect linking**
    - ▶ Surname averages are assigned through the observed father-child link.
    - ▶ This eliminates surname misspellings and name-standardization errors.
  2. **Whole Census**
    - ▶ Surname averages are computed using the full working-age male population in 1920.
    - ▶ No sampling of the surname distribution is imposed.
  3. **Full overlap** (Santavirta and Stuhler, 2024)
    - ▶ We keep only parent-child pairs where the father is included in the surname-average sample.
    - ▶ This avoids attenuation arising from non-overlapping surname means.

## Standard estimates in the benchmark

- **Empirical fact:** The surname estimator is much larger than the standard IGE

	(1) Parent-child	(2) Surname-based
Father OccScore	0.391*** (0.000)	
Surname Average Occscore		0.597*** (0.001)
Observations	4,463,879	4,463,879
R-squared	0.137	0.036

## Standard estimates in the benchmark

- **Empirical fact:** The surname estimator is much larger than the standard IGE

	(1) Parent-child	(2) Surname-based
Father OccScore	0.391*** (0.000)	
Surname Average Occscore		0.597*** (0.001)
Observations	4,463,879	4,463,879
R-squared	0.137	0.036

- **Why is that so? How should we interpret this?**

## A Stylized Framework: Multiplicity

- ▶ We model income  $y_{it}$  as the outcome of **multiple** underlying characteristics  $X_{it}^j$ , each affecting income with a *relevance* parameter  $\rho_j$ :

$$y_{it} = \sum_j \rho_j X_{it}^j + u_{it}$$

- ▶ Each characteristic is transmitted from parents to children with its own *persistence* parameter  $\lambda_j$ :

$$X_{it}^j = \lambda_j X_{i,t-1}^j + \varepsilon_{it}^j$$

- ▶ Key ingredients of the **simple** framework:
  - multiple transmission pathways  $X^j$  (education, location, skills, etc.)
  - heterogeneous intergenerational persistence  $\lambda_j$

## Estimators under the model

- ▶ **Key message:** intergenerational estimators identify *weighted averages* of the persistence parameters  $\lambda_j$ .

- ▶ **Parent–child estimator**

→ The parent–child correlation corresponds to:

$$\beta^{PC} = \frac{\sum_{j=1}^J \lambda_j \omega_j}{\sum_{j=1}^J \omega_j + \omega_u}, \quad \omega_j = \rho_j^2 V(X_{i,t-1}^j)$$

→ Weights reflect how much each channel contributes to income variation.

- ▶ **Surname-based estimator**

→ The surname estimator corresponds to a different weighting:

$$\beta^S = \frac{\sum_{j=1}^J \lambda_j \omega_j^S}{\sum_{j=1}^J \omega_j^S}$$

→ The crucial question: *what determines the weights  $\omega_j^S$ ?*

## Why Are Name-Based Estimates Large?

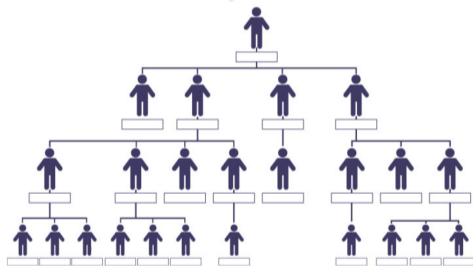
- ▶ Surname estimators place weight on traits that vary *across surname groups* or that are **shared within-group**: (Del Pizzo and Stuhler, 2025)

$$\omega_j^S = \rho_j^2 V\left(E[X_{i,t-1}^j \mid S_{i,t-1}]\right)$$

- ▶ What generates between-surname variation (or within-surname correlation)?
  - Members of a surname share some *common ancestor* in the past
  - Only traits that are **highly persistent** survive long enough to still differ across surname groups today
- ▶ Therefore, surname-based estimators place large weight on **very persistent traits**
  - Very useful to estimate Long-Run Persistence (e.g. historical settings)

# Why Surnames Capture Persistent Traits?

- ▶ Suppose two traits differ across families:
  - a **highly persistent** trait (e.g., geography)
  - a **low-persistence** trait (e.g., health)
- ▶ In early generations (e.g., second row)
  - **both** traits may still be shared among relatives
- ▶ In later generations (bottom row),
  - only the **persistent** trait remains shared across descendants
  - **weakly persistent** traits wash out



# Surname-based estimates: predictions and evidence

## ► Small surname groups

- recent common ancestor
- both *highly* and *weakly* persistent traits may still be shared

## ► Large surname groups

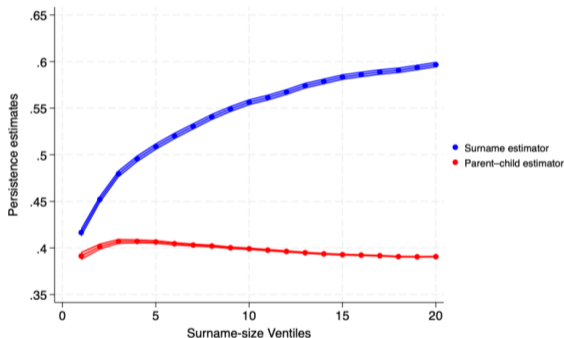
- distant common ancestor
- only *highly* persistent traits survive many generations

## ► Prediction

- surname-based persistence increases with *surname size*

► Non Smooth

► Descriptives



# The weighting structure of surname estimators

- ▶ In the model, the surname estimator increases because the weighting structure shifts across surname group size:
- ▶ Two key drivers:
  1. Idiosyncratic income variation is averaged out:  $\omega_u^S = 0$
  2. Traits with higher intergenerational persistence receive relatively larger weights:

$$\omega_j^S \uparrow \quad \text{for high } \lambda_j$$

- ▶ This raises an **empirical question**:
  - Which characteristics become more heavily weighted as surname size grows?
  - Can we recover this weighting structure directly from the data?

# Informational Content of Surnames (ICS) and $R^2$

- ▶ Güell et al. (2015) propose to use how well surnames predict economic outcomes as an indirect measure of intergenerational persistence (**Informational Content of Surnames**). [▶ Figure](#)

→ *Intuition:* The more surnames predict status the more society is rigid

- ▶ The estimator is essentially the  $R^2$  adjusted from regressing income onto surname indicators

$$ICS \approx \frac{V(\bar{y}_{it})}{V(y_{it})}$$

- ▶ The Weights are essentially a decomposition of the ICS

$$\hat{\omega}_j^S \approx \frac{V(\hat{y}_{it})}{V(\bar{y}_{it})}$$

## Empirical application: recovering the weights

**Goal:** Recover the model-implied weights  $\hat{\omega}_j^S$  from the data

**Empirical strategy:**

1. *Regress income on characteristic  $j$  and get fitted values*

$$y_{it}^j = E[y_{ist} \mid X_{ist}^j] = \rho_j X_{ist}^j, \quad y_{it}^{-j} = y_{ist} - y_{it}^j$$

2. *Average fitted values over surname groups*

$$\bar{y}_{ist}^g = E_n[y_{ist}^g \mid \text{surname}] = \rho_g \bar{X}_{ist}^g, \quad g \in \{j, -j\}$$

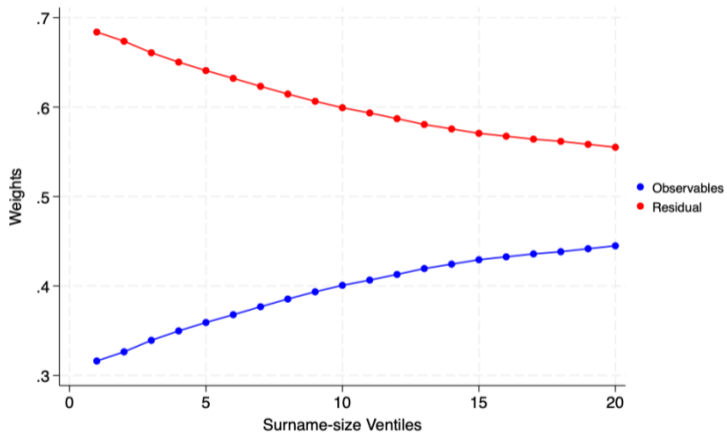
3. *Compute between-surname variance by surname size*

$$V(\bar{y}_{ist}^g \mid n_s < N) = \rho_g^2 V(\bar{f}_{ist}^g \mid n_s < N), \quad g \in \{j, -j\}$$

Go to Code!

**Go To Code!**

# Weights: Observables vs Unobservables



## Weights: normalization and comparison

- ▶ To compare the persistence of different characteristics, we examine how their **weights scale with surname group size**.
- ▶ **Key model implication**
  - A more persistent characteristic exhibits faster weight growth as surname groups expand:

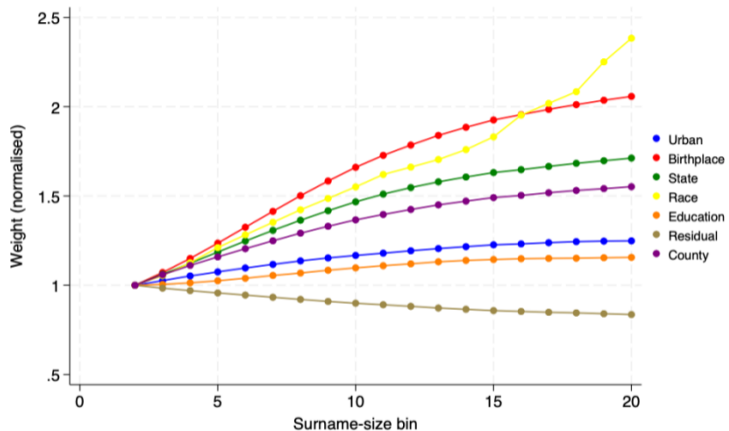
$$\frac{\omega_{j,N}}{\omega_{j,N_1}} \geq \frac{\omega_{-j,N}}{\omega_{-j,N_1}} \quad \text{if } \lambda_j \geq \lambda_{-j}.$$

- ▶ **Normalization: isolating persistence**

- Taking ratios across surname sizes differences out relevance parameters ( $\rho_j$ ), which do not vary with  $N$ .
- The resulting growth rate depends only on persistence:

$$\frac{\hat{\omega}_{j,N}}{\hat{\omega}_{j,N_1}} = \underbrace{\frac{V(\bar{f}_{jist} \mid n_s < N)}{V(\bar{f}_{jist} \mid n_s < N_1)}}_{= h(\lambda_j)} \times \frac{V(\bar{y}_{ist} \mid n_s < N_1)}{V(\bar{y}_{ist} \mid n_s < N)}.$$

# Weight Growth: Evidence



## Next step: relaxing assumptions one by one

- ▶ So far, results rely on an idealized setting with **perfect data**.
- ▶ In practice, surname-based estimators are particularly sensitive to **measurement error**, which is pervasive in historical and administrative data.
- ▶ We now relax the data assumptions *sequentially* to study how the surname estimator responds (Del Pizzo et al., 2025)
- ▶ In general, we introduce **measurement error**  $\Rightarrow$  attenuation bias.

Setting	Perfect linking	Whole Census	Full overlap
Benchmark	✓	✓	✓

# 1: Alternative Linking Strategies

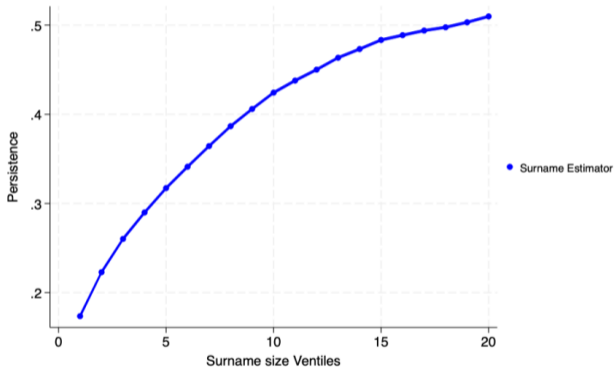
- ▶ In the census, surnames are frequently misspelled, especially rare ones
- ▶ We try different linking strategies:
  1. Surname averages on surnames from 1920 → Same surname father/child
  2. Surname averages on surnames from 1940 → Misspelling in surnames
  3. Link directly on surnames → Not perfectly linked sample

Setting	Perfect linking	Whole Census	Full overlap
Benchmark	✓	✓	✓
Alternative linking	✗	✓	✓

# 1: Results - Alternative Linking

## ► Main findings

- Measurement error generates attenuation bias.
- The effect is concentrated among **rare (small) surnames**.
- This steepens the estimator's gradient with respect to surname size.



## 2: Sampling The Census

- ▶ There are cases in which we have samples and not the whole census. e.g. 1% representative sample US
- ▶ Strategy:
  1. Take a random sample of 5% of the original data
  2. Compute the surname average *in the sample*
  3. Estimate the surname estimator with the surname average computed in the census and in the sample

Setting	Perfect linking	Whole Census	Full overlap
Benchmark	✓	✓	✓
5% Census sample	✓	✗	✓

## 2: Results - Sampling The Census

### ► Main findings

- Computing surname averages *within the estimation sample* pulls the surname estimator toward the parent-child correlation.
- *Intuition:* with small surname groups and full overlap, surname means increasingly approach the father's outcome.
- In the model, sampling noise makes surname-based weights noisier, attenuating the estimate.

	(1) Parent-child	(2) Surname-Census	(3) Surname-Sample
Father OccScore	0.391*** (0.002)		
Surname Average (Census)		0.592*** (0.007)	
Surname Average (Sample)			0.435*** (0.004)
Observations	223,194	223,194	223,194
R-squared	0.138	0.035	0.056

### 3: Limited Overlap

- ▶ Draw independent random 5% samples from the 1920 and 1940 censuses
- ▶ Compute surname-level averages using the 1920 sample and merge them into the 1940 sample
- ▶ This design minimizes **overlap**: a father appears in the surname average with only 5% probability

Setting	Perfect linking	Whole Census	Full overlap
Benchmark	✓	✓	✓
Reduced overlap	✓	✗	✗

### 3: Results - Limited Overlap

#### ► Main findings

1. With limited overlap, the surname estimator falls *below* the parent-child correlation
2. In the model, reduced overlap shrinks the numerator while leaving the denominator unchanged: → Attenuation toward zero

	(1) Parent-child	(2) Surname-Census	(3) Surname-Sample	(4) Surname-No Overlap
Father OccScore	0.391*** (0.002)			
Surname Average (Census)		0.592*** (0.007)		
Surname Average (Sample)			0.435*** (0.004)	
Surname Average (Limited overlap)				0.286*** (0.006)
Observations	223,194	223,194	223,194	197,485
R-squared	0.138	0.035	0.056	0.011

# Hands-on Replication Using the Public 1% Sample

- ▶ In this lab session, participants will replicate the key empirical patterns from the lecture:
  1. Verify that surname-based and parent-child (direct) estimators deliver similar results in a small sample
  2. Show concretely how **imperfect linking** (e.g., misspellings) leads to attenuation and lower estimated persistence
  3. Demonstrate that constructing surname averages with **limited overlap** across generations produces severe downward bias.
- ▶ These exercises illustrate, step by step, how data imperfections mechanically affect surname estimators.

## From weights to geography

- ▶ So far, we showed that surname estimators overweight **highly persistent characteristics**.
- ▶ Geography is a major driver (Chetty et al., 2014):
  - highly persistent over generations (hence, correlated with surnames)
  - economically meaningful
- ▶ We now use geography as a **case study** (Del Pizzo et al., 2025)
  1. to show how **conditioning** on a persistent group-level trait affects surname-based and parent-child estimates,
  2. and to assess whether surname estimators can be used to **compare** persistence *across* geographic areas.

## Conditioning on Geography

- ▶ Surname-based estimators place relatively more weight on **geographic persistence** than parent-child estimates.
- ▶ Conditioning on increasingly granular geographic controls brings the weight to zero:  $\omega_{geo} \rightarrow 0$ :
  - Since  $\omega_{geo}^S > \omega_{geo}^{PC}$ , the surname estimator and the parent-child estimate mechanically converge.
- ▶ **Caveat:** This is **NOT** a bias correction → it is a **different estimand**.
  - The remaining persistence measures transmission *net of* geography, not overall intergenerational persistence.

## Estimates Conditional on Geography

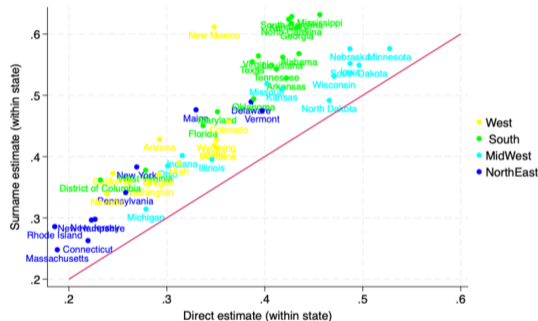
	(1) None	(2) Macro FE	(3) State FE	(4) County FE
<b>Panel A. Surname-based estimator</b>				
Surname-based estimator	0.597*** (0.001)	0.526*** (0.001)	0.446*** (0.001)	0.334*** (0.001)
FE level	None	Macro FE	State FE	County FE
Observations	4,463,879	4,463,879	4,463,879	4,463,879
R-squared	0.036	0.066	0.103	0.172
<b>Panel B. Parent-child estimator</b>				
Parent-child estimator	0.391*** (0.000)	0.365*** (0.000)	0.342*** (0.001)	0.286*** (0.001)
FE level	None	Macro FE	State FE	County FE
Observations	4,463,879	4,463,879	4,463,879	4,463,879
R-squared	0.137	0.145	0.160	0.198

## Comparing Persistence Across Areas

- ▶ A common goal in the mobility literature is to compare intergenerational persistence across geographic areas (Chetty et al., 2014).
- ▶ Even if surname-based estimators differ in levels from parent-child estimates, they may still capture a similar **ranking** of areas.
  - *Empirically* surname-based and parent-child estimates display very similar cross-area patterns.
- ▶ **However, discrepancies can arise for two conceptual reasons:**
  1. **Surname composition:** areas with more common surnames mechanically exhibit higher surname-based persistence.
  2. **Heterogeneous relevance of persistent traits:** some highly persistent characteristics (e.g. race) matter more in certain regions (e.g. the U.S. South).

# Ranking Persistence Across Areas

- ▶ For each state (or county), we estimate:
  - a surname-based persistence measure,
  - a parent-child persistence measure.
- ▶ **Main result:**
  - persistence rankings across areas are highly similar,
  - the South is the least mobile,
  - the North-East is the most mobile.



## Conclusion

- ▶ Surname-based estimators place large weights on **highly persistent traits**, such as geography or other group-level characteristics.
  - They are useful for quantifying **long-run persistence**
- ▶ They are, however, sensitive to data imperfections:
  - sampling of the surname distribution,
  - imperfect linking and name measurement error,
- ▶ Controlling for persistent geography, surname-based and parent-child estimates yield **very similar** rankings across areas.

## References I

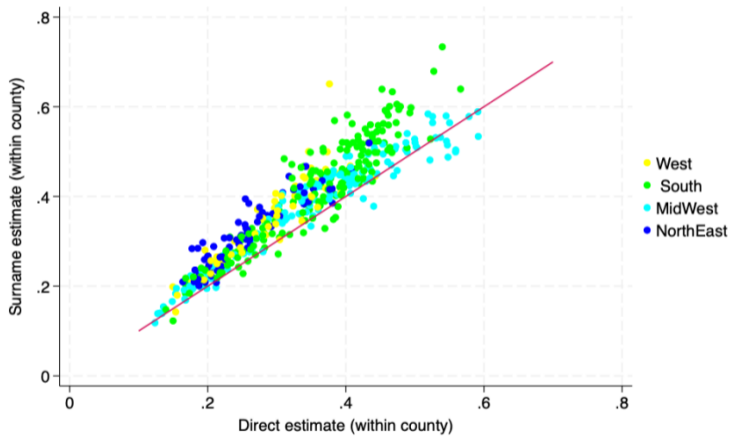
- Guglielmo Barone and Sauro Mocetti. Intergenerational mobility in the very long run: Florence 1427–2011. *The Review of Economic Studies*, 88(4):1863–1891, 2021.
- Marianna Belloc, Francesco Drago, Mattia Fochesato, and Roberto Galbiati. Multigenerational transmission of wealth: Florence, 1403–1480. *American Economic Journal: Applied Economics*, 16(2): 99–129, 2024.
- Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014.
- Gregory Clark. The son also rises. In *The Son Also Rises*. Princeton University Press, 2014.
- M Dolores Collado, Ignacio Ortuño-Ortín, and Andrés Romeu. Long-run intergenerational social mobility and the distribution of surnames. *manuscript, Universidad de Alicante*, 2012.
- Andrea Del Pizzo and Jan Stuhler. Multiplicity in intergenerational transmission: Evidence from surnames. *Working Paper, Universidad Carlos III Madrid*, 2025.
- Andrea Del Pizzo, Martin Nybom, and Jan Stuhler. Indirect estimators of intergenerational mobility. *Working Paper*, 2025.
- Maia Güell, José V Rodríguez Mora, and Christopher I Telmer. The informational content of surnames, the evolution of intergenerational mobility, and assortative mating. *The Review of Economic Studies*, 82(2):693–735, 2015.

## References II

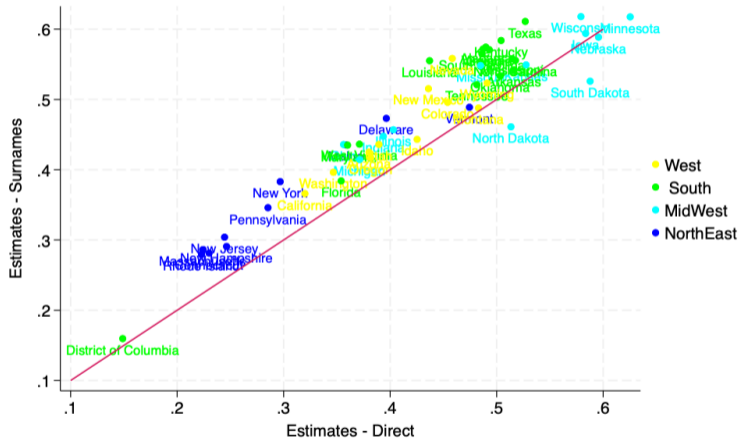
Claudia Olivetti and M Daniele Paserman. In the name of the son (and the daughter): Intergenerational mobility in the united states, 1850–1940. *American Economic Review*, 105(8):2695–2724, 2015.

Torsten Santavirta and Jan Stuhler. Name-Based Estimators of Intergenerational Mobility. *The Economic Journal*, 134(663):2982–3016, 05 2024.

# Counties



# State - Only rare surnames



# Surname Group Size

- ▶ We define the size of surname group  $s$  at time  $t$  as:

$$n_{s,t} = n_{s,\tau_s} \cdot g_s^{t-\tau_s}$$

where:

- $t - \tau_s$  = time since the MRCA (TMRCA)
  - $g_s$  = net fertility rate specific to surname  $s$
  - $n_{s,\tau_s}$  = number of Most Recent Common Ancestors (MRCAs)
- ▶ **Key Insight:** If  $g_s > 1$ , then longer TMRCA mechanically implies a larger current group size.
  - ▶ **Assumption:** Other elements affecting group size are not strong enough to reverse this positive relationship.

- ▶ The **Informational Content of Surnames** (ICS)  
(Güell et al., 2015)

$$ICS = R_{Sur}^2 - R_{Fakesur}^2$$

- ▶ *Intuition:* It represents the share of the outcome variable explained by surnames

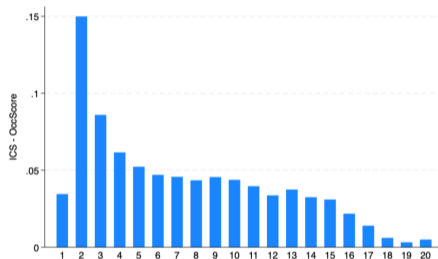
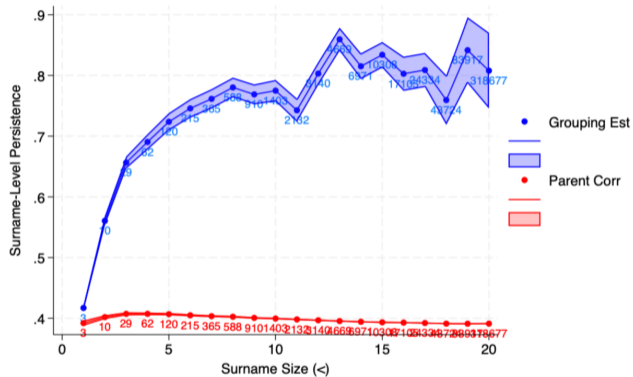


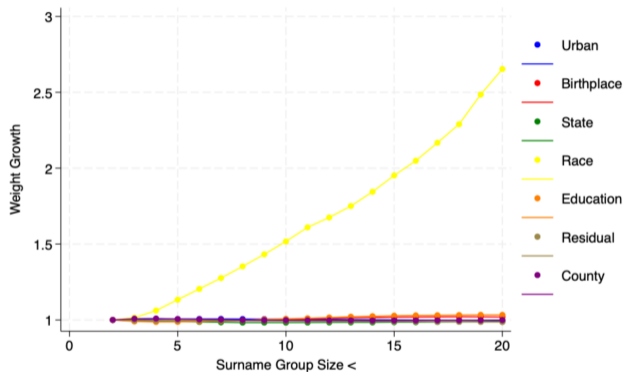
Figure 1: ICS across ventiles of surname group size

# Grouping estimator and Parent-Child Correlation - No Smoothing



► Back

# Weights' Growth - Parent Child Correlation



► Back

# Sample and summary statistics

Table 1: Summary Statistics by Quintiles of Surname Group Size

	Quintiles of Surname Size					2*Total
	1	2	3	4	5	
N	895,044	890,533	892,833	894,598	890,726	4,463,734
Child Occupational Score	3.168	3.146	3.140	3.138	3.106	3.140
Years of Schooling	9.438	9.524	9.557	9.549	9.294	9.473
<i>Race/Ethnicity</i>						
Non-White	(2.6%)	(3.9%)	(5.1%)	(6.4%)	(9.7%)	(5.5%)
White Non-Hispanic	(97.4%)	(96.1%)	(94.9%)	(93.6%)	(90.3%)	(94.5%)
<i>Region</i>						
Northeast	(33.0%)	(27.6%)	(24.8%)	(24.3%)	(20.0%)	(25.9%)
Midwest	(42.0%)	(37.1%)	(33.2%)	(31.2%)	(30.7%)	(34.8%)
South	(15.6%)	(25.1%)	(31.1%)	(33.4%)	(38.2%)	(28.7%)
West	(9.5%)	(10.3%)	(10.8%)	(11.1%)	(11.0%)	(10.5%)
<i>Urban/Rural Status</i>						
Rural	(39.0%)	(45.9%)	(48.3%)	(48.1%)	(51.3%)	(46.5%)
Urban	(61.0%)	(54.1%)	(51.7%)	(51.9%)	(48.7%)	(53.5%)