CIVIC DIGITAL FELLOWSHIP

Revenue-Based County Business Patterns

Sikata Sengupta and Drishaan Jain

Supervised by Javier Miranda and Anne Russell

Economy-Wide Statistics Division

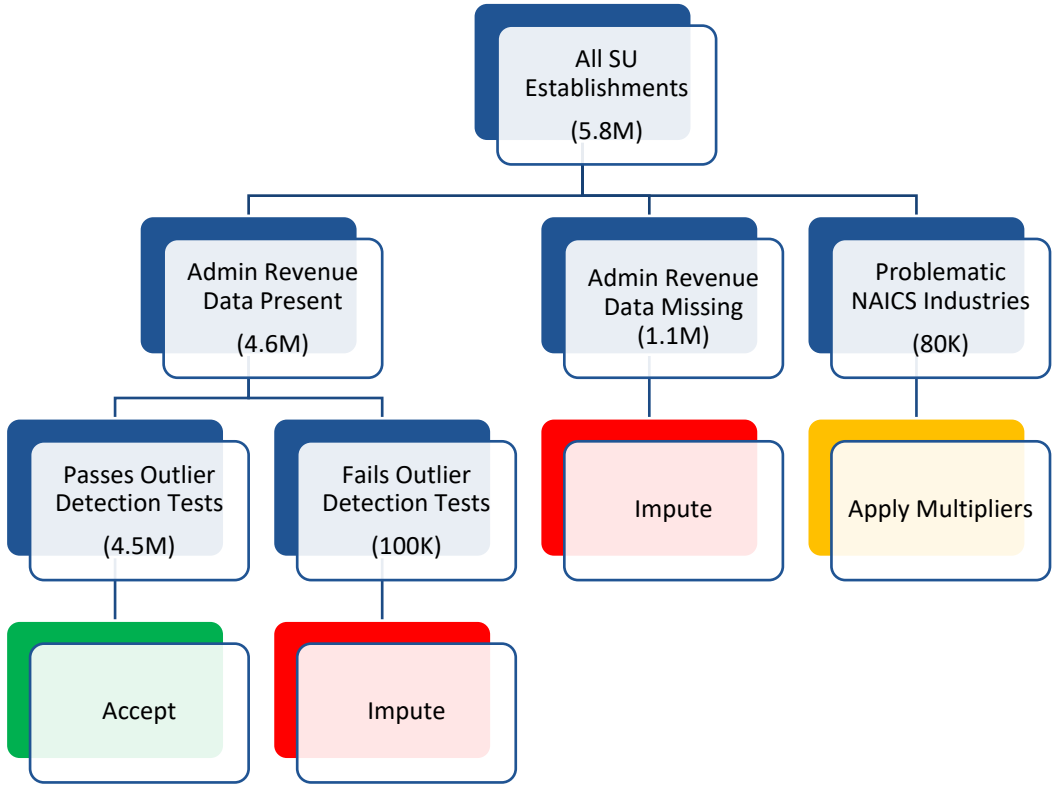
AGENDA

- ❑ Recap of Single Unit Imputation
- ❑ Multi Units: Approach and Challenges
- ❑ Multi Units: Imputation Results
- ❑ Deep-Dive/Discrepancy Analysis
- ❑ Goals and Next Steps

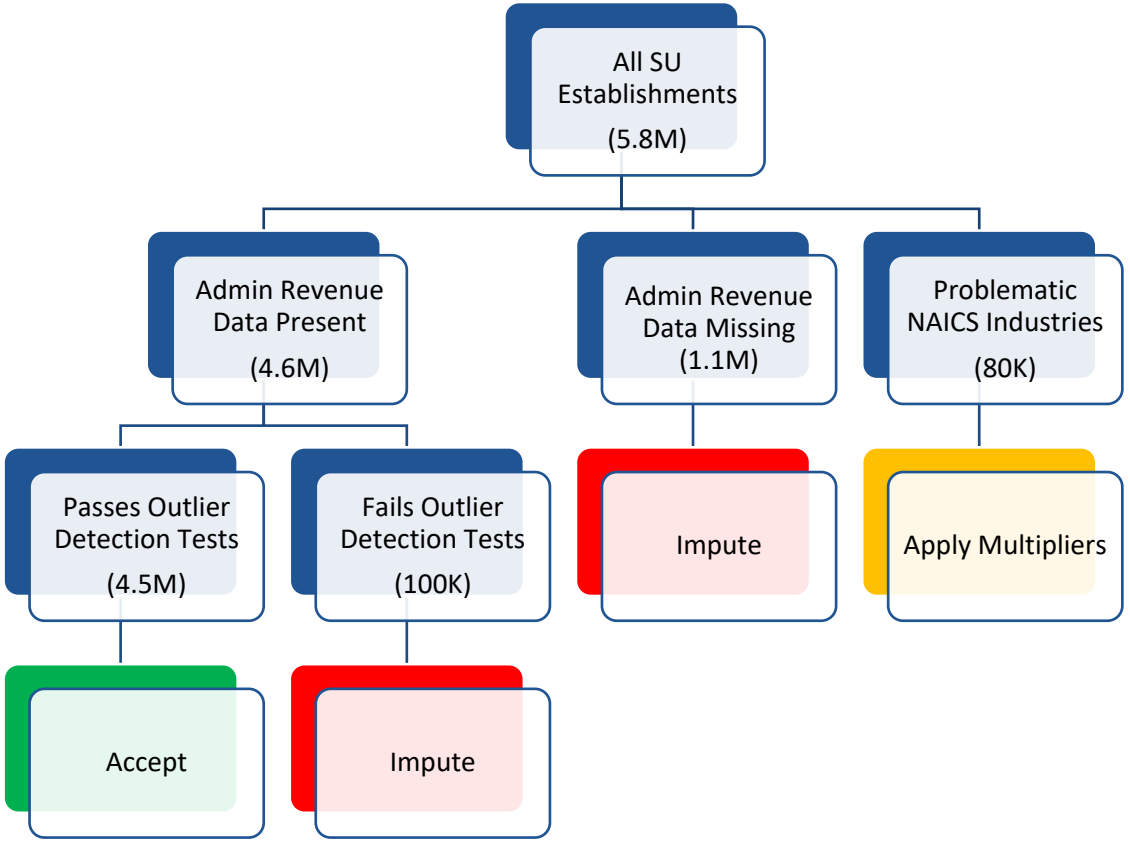
SINGLE UNITS: RECAP & SUMMARY

SU APPROACH

- Inclusion of revenue data in CBP SU realm
  - Broken into groups to accept vs. impute: acceptable, outliers, missing data, problematic NAICS industries
- Administrative Data: revenue data and revenue quality flag information integrated
- Economic Census Data: revenue data solely used for comparison/verification purposes with the administrative data; no imputations will come from these data



FINAL SU RESULTS



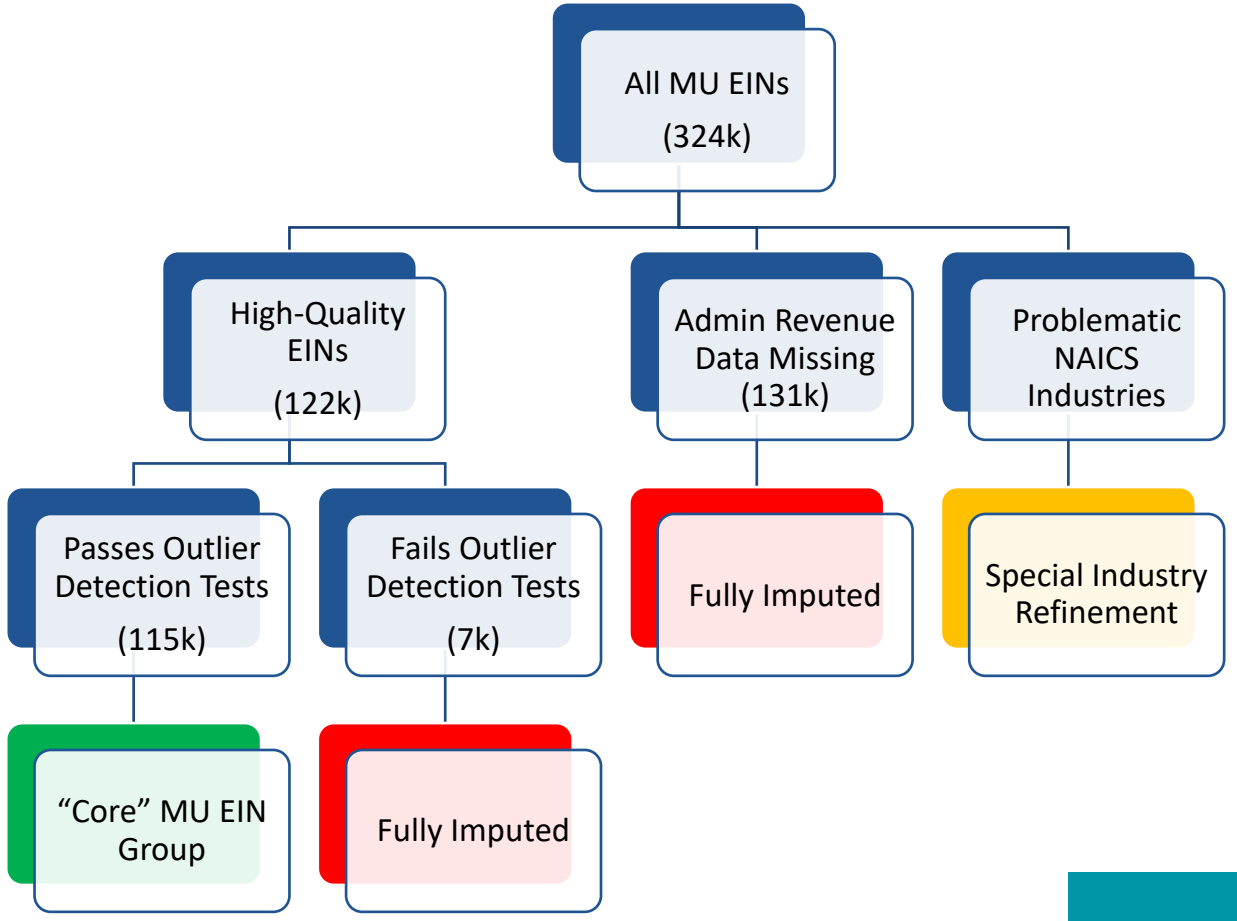
Comparison of Administrative Revenue Data vs. Economic Census Revenue Data

	Before	After
Total Revenue % Difference (Admin vs. EC)	40% higher	1.5% lower
Average Establishment-Level $ Difference (Admin vs. EC)	$681K higher	$25K lower

MULTI UNITS: APPROACH & CHALLENGES

MU APPROACH #1: EIN-Based

- Subset of MU EINs meeting certain quality rules and passing all outlier detection tests
  - Approx. 115k / 324k EINs (35.5%)
- Use subset to compute revenue-payroll ratios at EIN level for each 4-digit industry
- Apply ratios to impute individual establishment-level CBP



APPROACH #1 CHALLENGES

- MU EIN subset skewed towards smaller MU firms – not representative of all MUs
- Revenue-payroll ratios created based on entire EIN made up of single industry (different EIN profiles) → not much power
  - No power to interact with firm size
- Loses state-level accuracy as EINs can span establishments across multiple states

Firm Size	EINs in Core Group	Total EINs
Firms with less than 5 establishments	96.6k (58.1%)	166.2k
Firms with 5 to 9 establishments	15.6k (34.1%)	45.9k
Firms with 10 to 19 establishments	5.7k (19.4%)	29.4k
Firms with 20 to 49 establishments	2.9k (10.8%)	27.3k
Firms with 50 to 99 establishments	750 (5.2%)	14.3k
Firms with 100 to 249 establishments	214 (1.4%)	15k
Firms with 250 to 499 establishments	27 (0.33%)	8.2k
Firms with 500 to 999 establishments	13 (0.24%)	5.4k
Firms with 1,000 establishments or more	3 (0.02%)	12.1k
TOTAL	**121.9k (37.63%)**	**324k**

MU APPROACH #2: Establishment Based

MOTIVATION: Estimate productivity relationship between SU and MU in the Economic Census, apply to SU Admin data to predict MU revenue

STEP 1: Comparing the MU and SU revenue-payroll ratios within the 2017 Economic Census

Approach

- ❑ Limited to EC respondent SU and MU establishments
- ❑ Aggregate revenue and payroll to 4-digit industry separately between SU and MU
- ❑ Generate revenue-payroll ratios for SU and MU in each industry

Industries with SU ratio > MU ratio

- ❑ 2111: Oil and Gas Extraction
- ❑ 5511: Management of Companies and Enterprises
- ❑ 5179: Other Telecommunications
- ❑ 4853: Taxi and Limousine Service

Conclusions

- ❑ In general, MU firms tend to have higher revenue-payroll ratios (more productive)
- ❑ For industries where SU and MU ratios line up closely, there are usually more SU establishments than MUs
  - ❑ Consider size of SU vs. MU (ex: comparing large SUs to small MUs)
- ❑ Majority of industries have > 10% difference between ratios
- ❑ Comparison to approach 1 favorable for most industries
  - ❑ Challenges: small cell size, selected industries

MU APPROACH #2: Establishment Based

Step 2: Comparing the MU revenue-payroll ratios between the Economic Census, core MU EIN group with no outliers, and adjusted SU ratios.

Approach

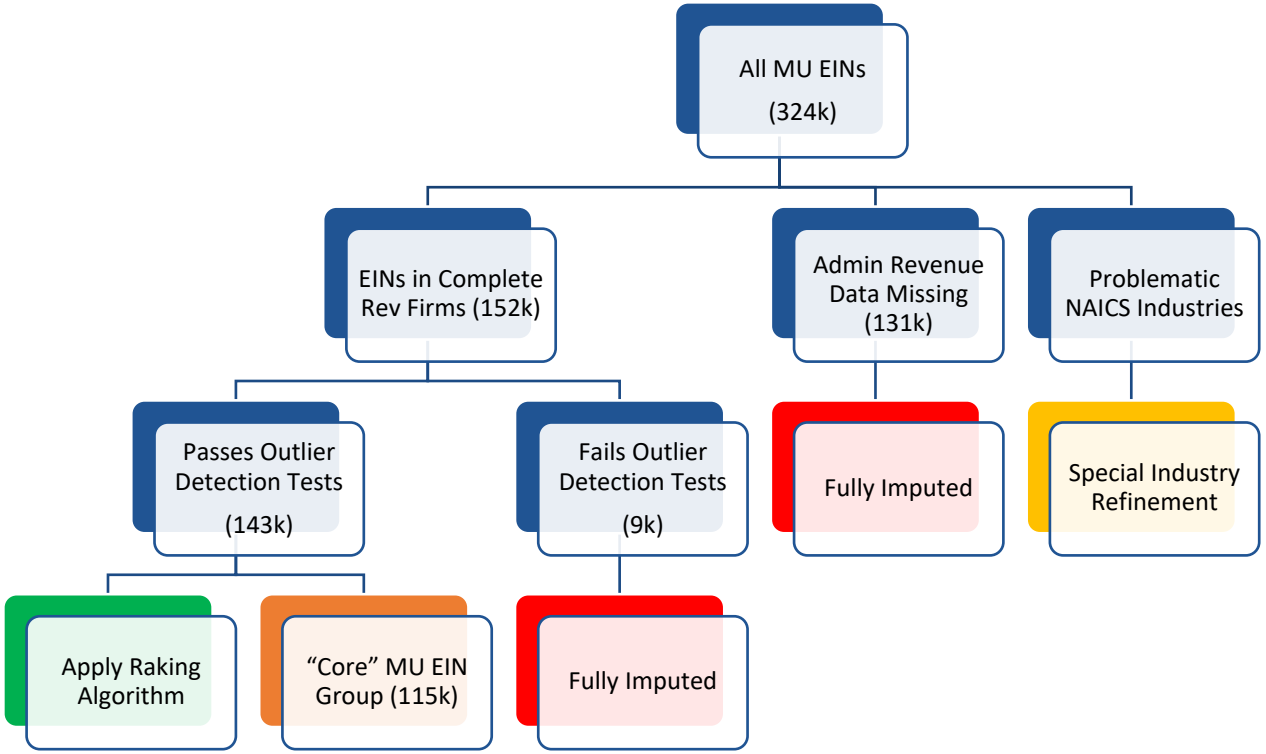
- ❑ Calculate revenue-payroll ratios using core MU EIN group with no outliers
- ❑ Use ratio between Census MU revenue-payroll ratios to Admin SU revenue-payroll ratios to develop adjustment factor
- ❑ Apply adjustment multiplier to final SU revenue-payroll ratios from 2017 analysis
- ❑ Compare EC MU ratio, core MU EIN ratio, and adjusted final SU ratio

Conclusions & Challenges

- ❑ Using SU adjusted ratios: multiplier will stay constant between Census years but SU ratios change every year
  - ❑ Affected by industry shocks, fundamental changes in productivity relationship between SU and MU
  - ❑ **Test:** how stable are the SU adjustment multipliers? Compare between 2012 and 2017 EC
- ❑ Scope differences of industries only covered by SU or MU, discrepancy of Admin vs. Economic Census
- ❑ Must use MU Admin data and auxiliary data to monitor and adjust for drift in the SU/MU productivity relationship at the industry level

MU: RAKING ALGORITHM

- After developing establishment-level imputed revenue data → aggregate to firm level in CBP
- Calculate percentage that each individual establishment contributes to total firm revenue
- For EINs belonging to firms with complete Admin revenue and passing outlier detection, aggregate EIN revenue to firm-level
  - 143k / 324k (44% of EINs, 30% of establishments)
- Apportion Admin firm level revenue to CBP establishments based on calculated percentages



MULTI UNITS: IMPUTATION RESULTS

IMPUTATION ROUND 1

CBP MU	R1 Imputation
Total Revenue % Difference (Imputations vs. EC)	1.25% higher
Average Establishment-Level $ Difference (Imputations vs. EC)	$169k higher

Approach & Conclusions

- ❑ Ratio factors: NAICS 4-digit
- ❑ Every individual establishment imputed
  - ❑ Regardless of core group, outlier detection, EIN/firm membership
  - ❑ Only reliant upon establishment 4-digit industry
- ❑ Good performance somewhat contrived due to creation of adjustment factors

IMPUTATION ROUND 2

CBP MU	R1 Imputation	R2 Imputation
Total Revenue % Difference (Imputations vs. EC)	1.25% higher	1.47% higher
Average Establishment-Level $ Difference (Imputations vs. EC)	$169k higher	$199k higher

Problematic Industries

- ❑ 4931: Warehousing and Storage
- ❑ 5242: Agencies, Brokerages, and Other Insurance Related Activities
- ❑ 4851: Urban Transit Systems

Approach & Conclusions

- ❑ Ratio factors: NAICS 4-digit, firm size
  - ❑ Four firm size classes (<5, 5-49, 50-249, 250+ establishments)
- ❑ Discrepancies dominated by problematic industry + size class combinations that negatively affect performance
- ❑ Reality: non-problematic industries achieve higher accuracy by incorporation of firm size
  - ❑ Improvement in approximately 86% of industry + size combinations resulting in less than 10% discrepancy

APPLYING RAKING

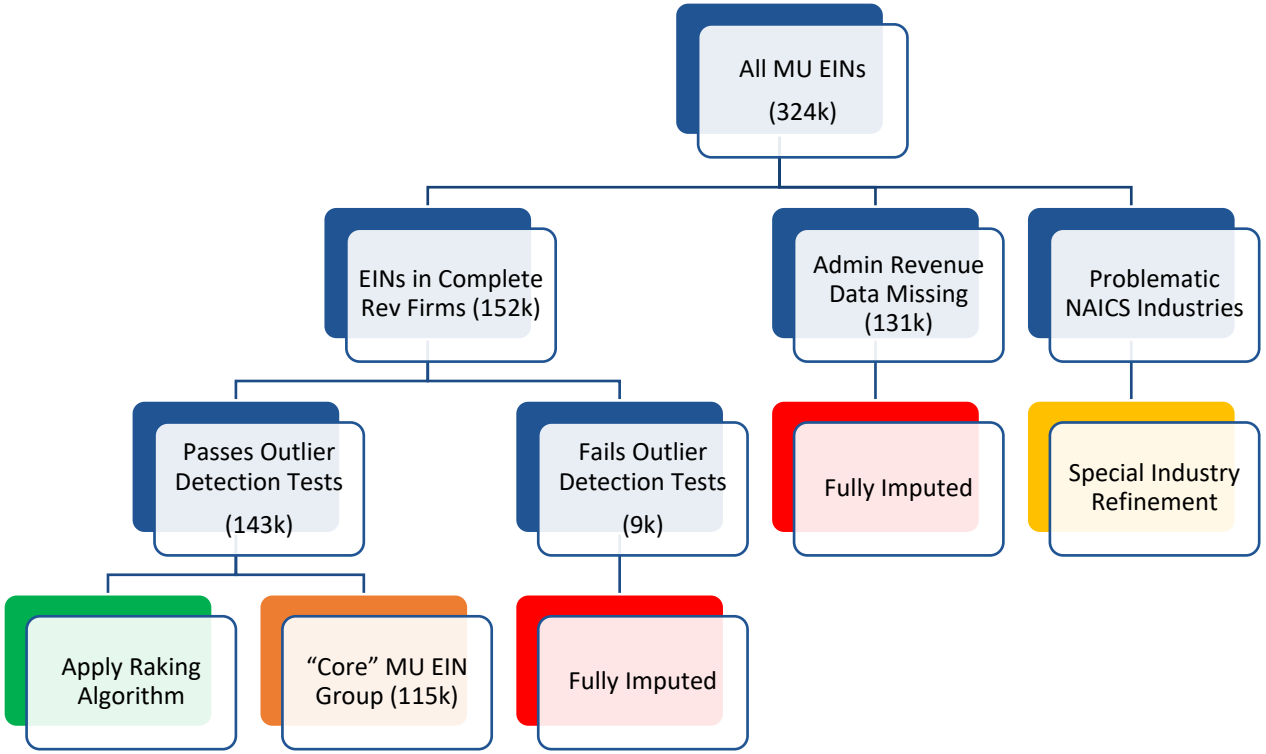
CBP MU	R1 Imputation	R2 Imputation	Raking
Total Revenue % Difference (Imputations vs. EC)	1.25% higher	1.47% higher	1.17% lower
Average Establishment-Level $ Difference (Imputations vs. EC)	$169k higher	$199k higher	$157k lower

Approach & Conclusions

- ❑ Applying raking algorithm using second round imputation results
- ❑ Only applicable to EINs within raking group subset
- ❑ "Best revenue" value: raked revenue if possible, imputed revenue accepted otherwise
- ❑ Raking can be applied with any imputation model to improve accuracy
  - ❑ Ex: in addition to future imputation model including geography

MULTI UNITS: NEXT STEPS

- Additional factors within imputation model, more advanced approaches
  - Namely: geography (state-level)
  - Regression, CART models
    - Firm/establishment age/county
- Special processing and refinement of problematic and out-of-scope industries
- Improvements
  - Using results of raking/approach #1 to recalculate adjustment factors from year to year, between Census
  - Including Annual Survey data to rake for larger MU firms



DEEP-DIVE

Scope

- **Scope Differences**
  - Need Apples to Apples Comparison
  - Outlying Territories
    - keep out for later analysis
- **SU vs. MU**
  - NAICS 5211

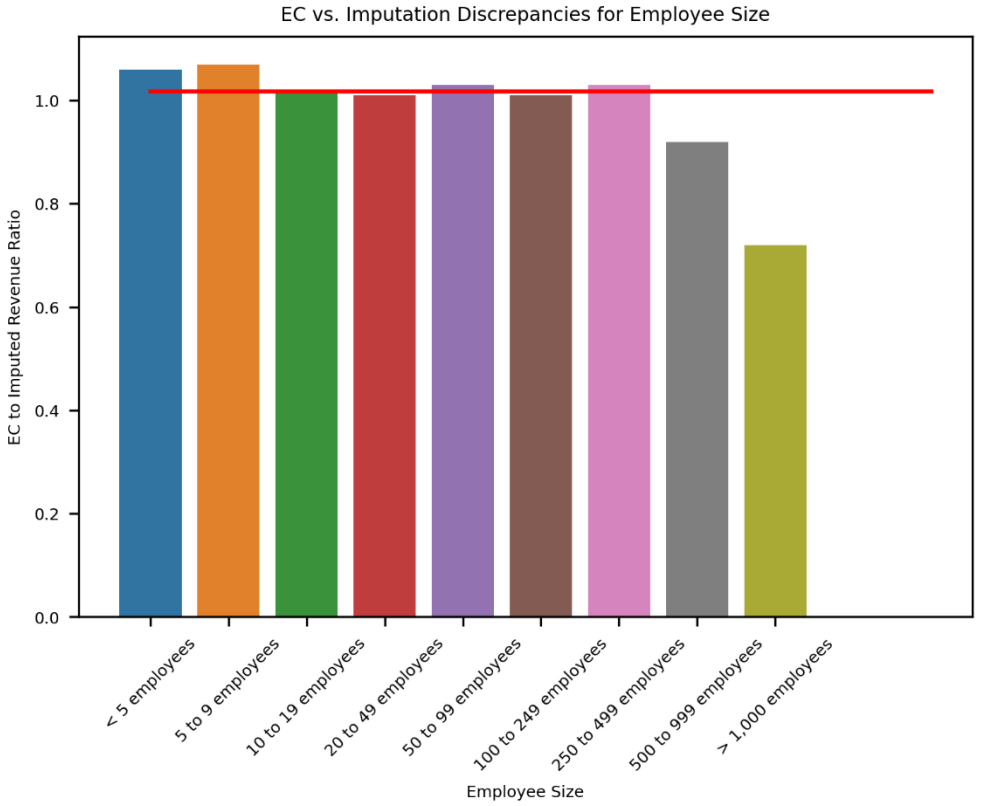
- **In-scope to CBP (not in EC)**
  - NAICS 113 Forestry and Logging
  - NAICS 114 Fishing, Hunting and Trapping
  - NAICS 115 Support Activities for Agriculture and Forestry
  - NAICS 525 Funds, Trusts, and Other Financial Vehicles
  - NAICS 52591 Open-End Investment Funds
  - NAICS 52593 Real Estate Investment Trusts
  - NAICS 52599 Other Financial Vehicles
  - NAICS 6111 Elementary & Secondary Schools (private schools only)
  - NAICS 6112 Junior Colleges (private schools only)
  - NAICS 6113 Colleges, Universities, and Professional Schools
  - (private schools only)
  - NAICS 8131 Religious Organizations
  - NAICS 813930 Labor Unions and Similar Labor Organizations
  - NAICS 813940 Political Organizations

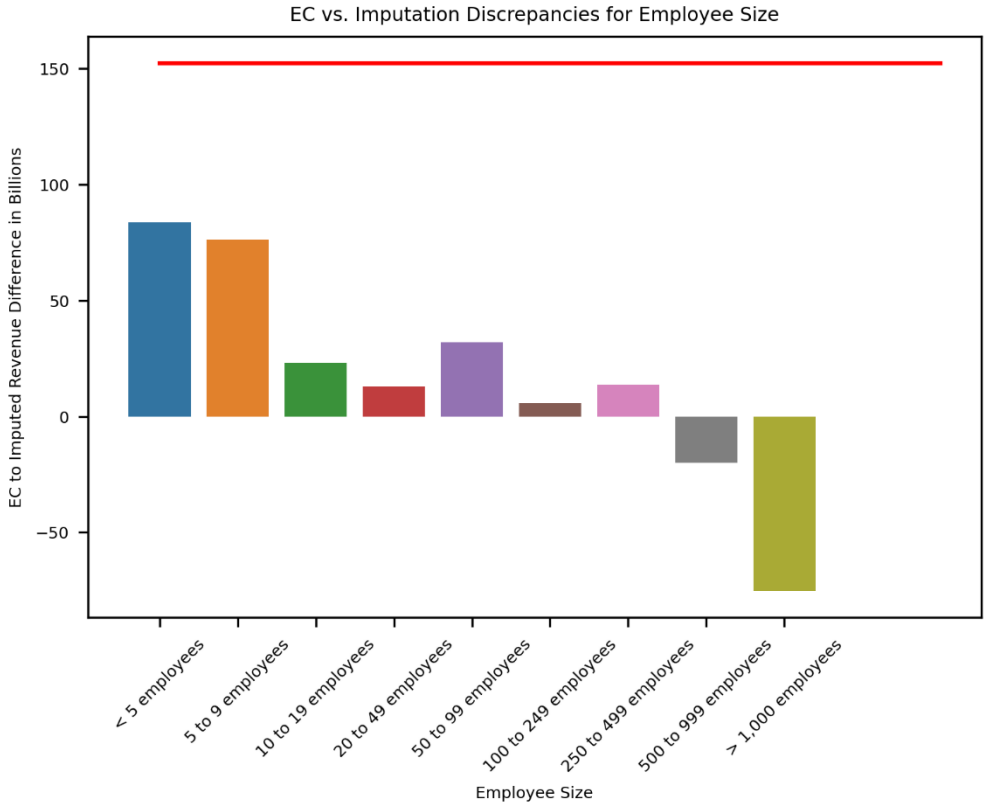
Analysis

- **Discrepancies with 2017 EC**
  - Stakeholders, Fully-Saturated SU
  - State
  - Sector
  - Size
  - Industry
- **Method of Comparison**
  - Ratios
  - Differences

SIZE-SU

- ❑ Larger size classes, over-estimate
- ❑ Perform quite well for smaller estabs

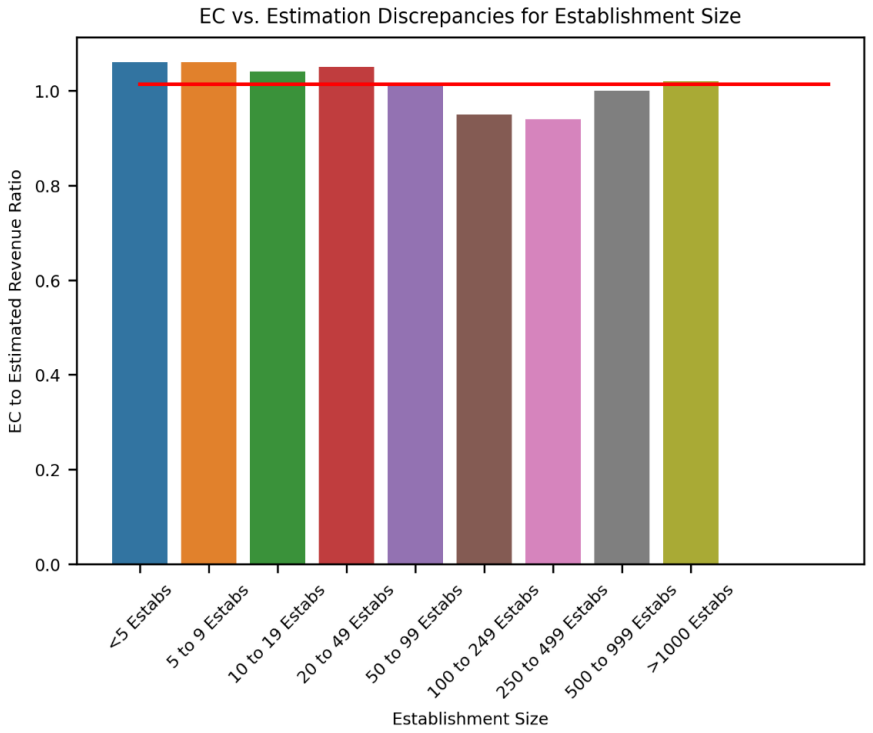


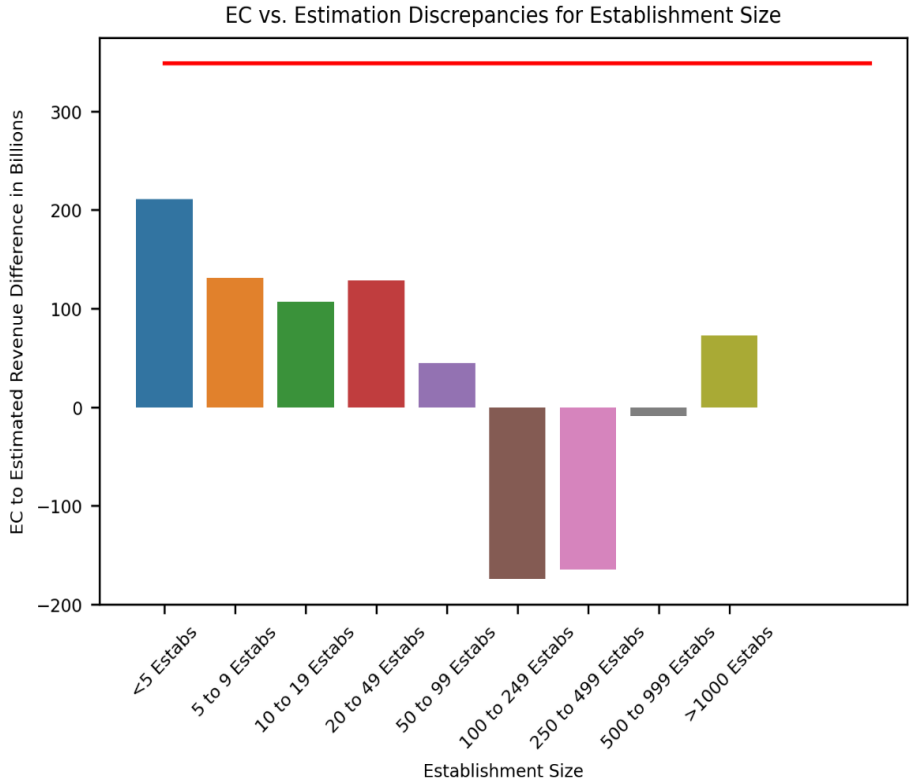


EW ratio: 1.02
EW diff: ~150 billion

SIZE-MU

- ❑ Mid-Size Firms over-estimate
- ❑ Perform quite well for larger firms

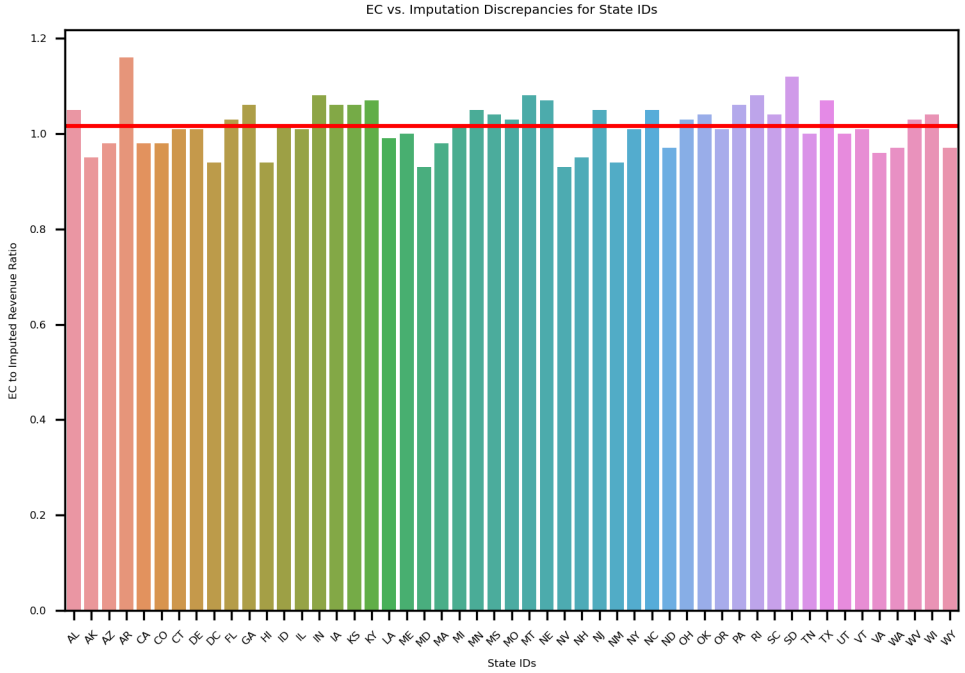


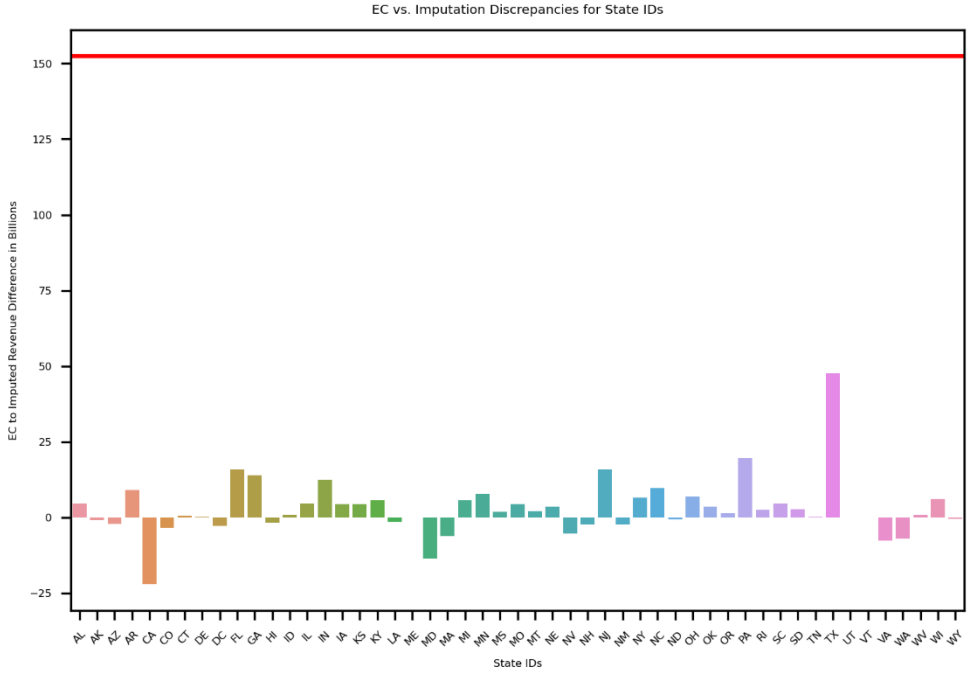


EW ratio: 1.01
EW diff: ~350 billion

STATES-SU

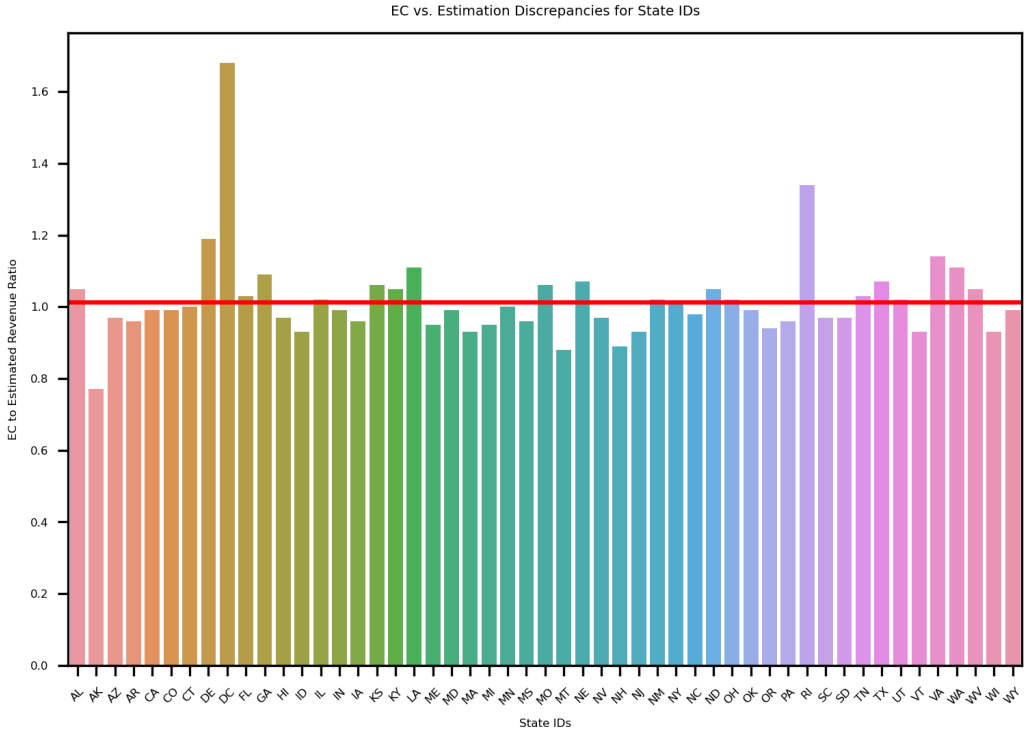
- ❑ Good on aggregate
- ❑ Larger discrepancies in larger states
- ❑ Territories, expected behavior

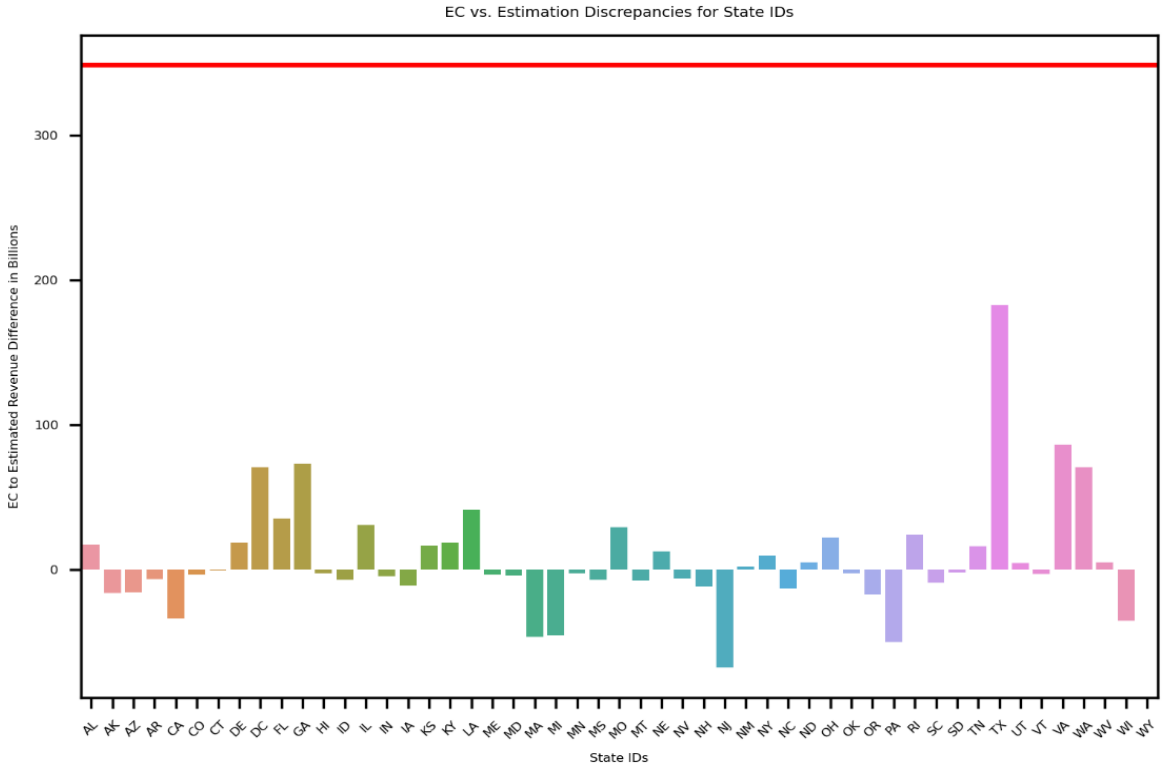




STATES-MU

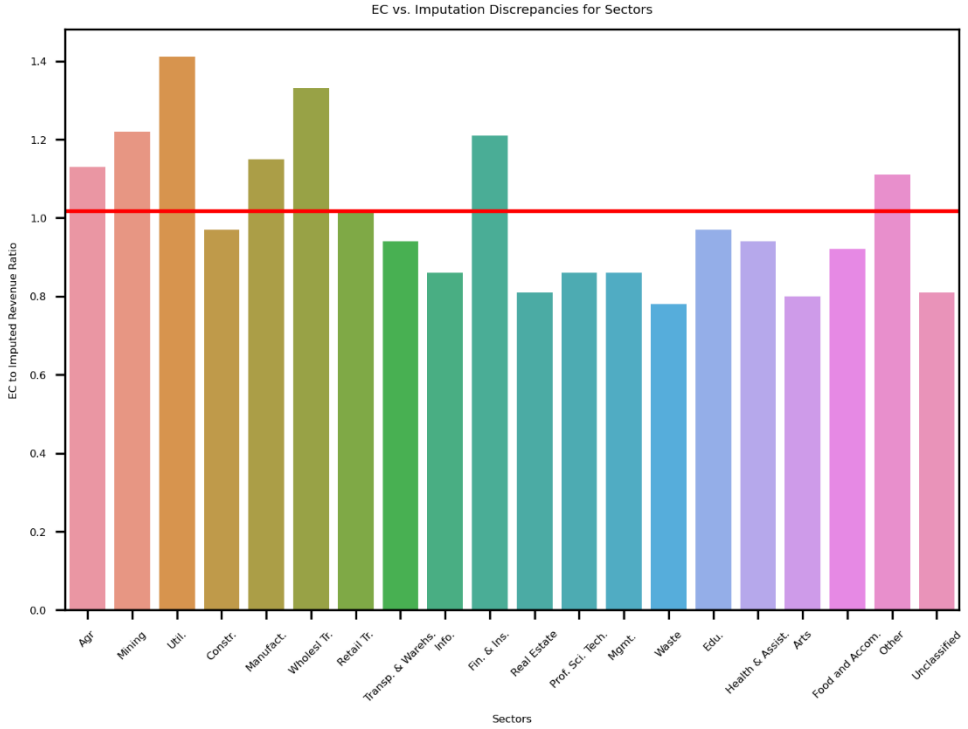
- Good on aggregate
- Larger discrepancies in larger states
- Ratios of Smaller Places: RI and DC
  - 4247=petroleum, 5222=nondepos. cred.

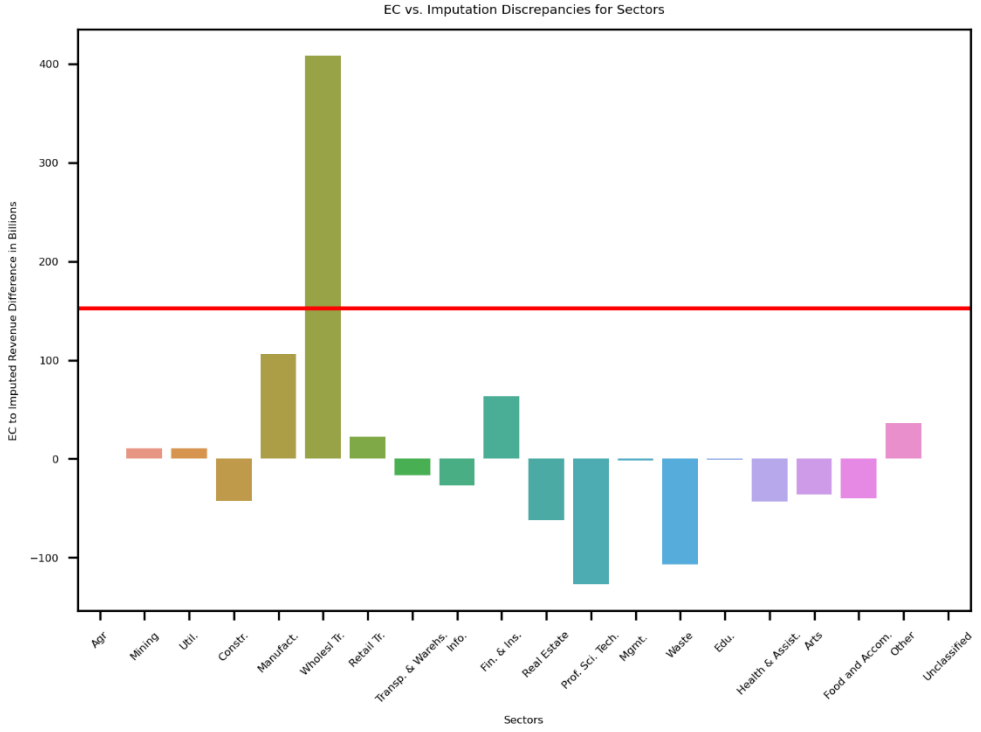




SECTOR-SU

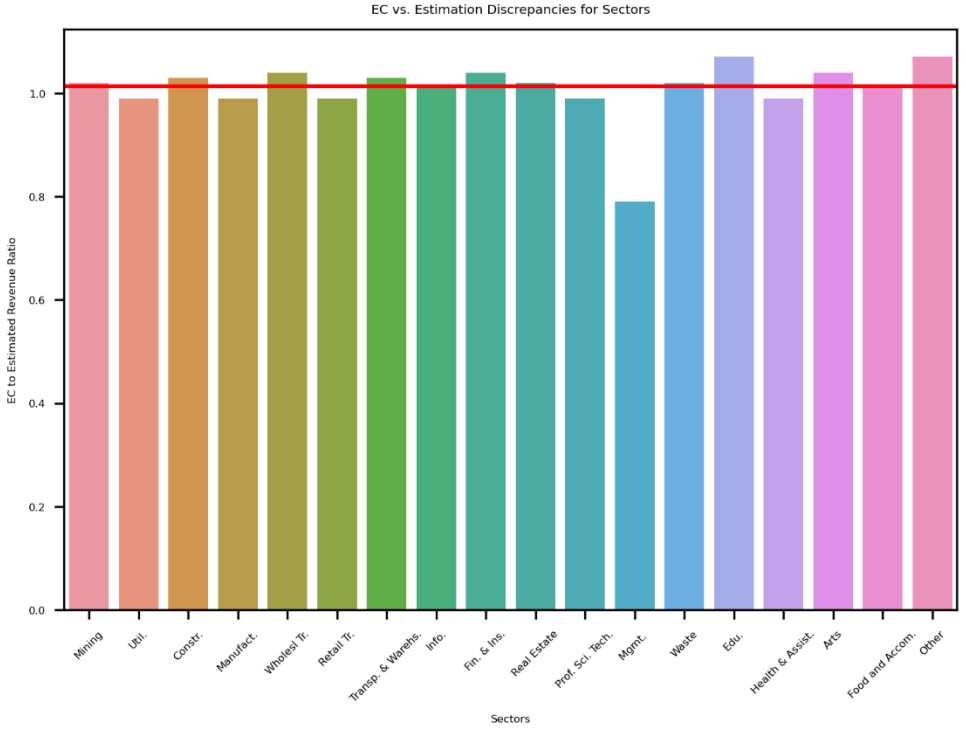
- More variation:
  - Wholesale Trade, Util, Manufact
- Differences, Wholesale Trade driver

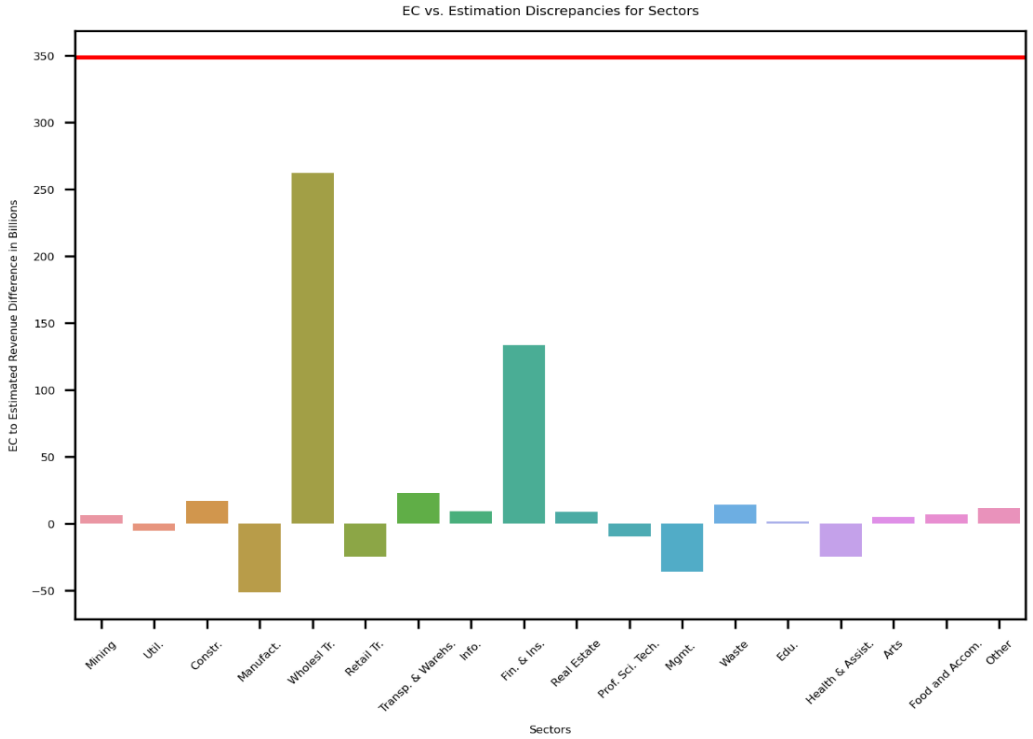




SECTOR-MU

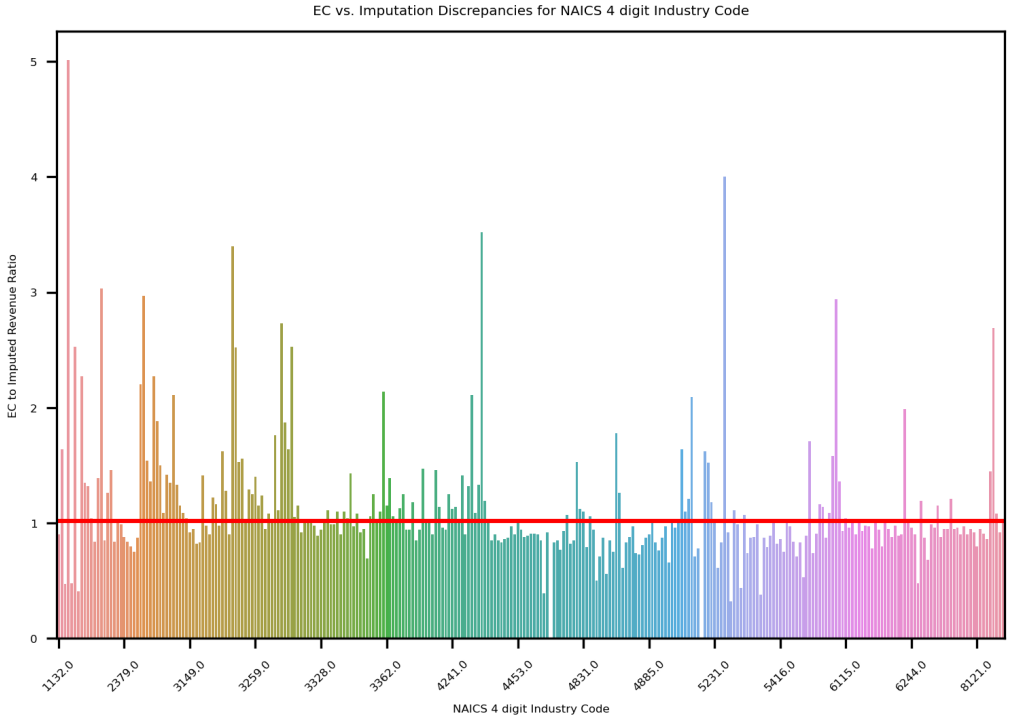
- Quite well:
  - Wholesale Trade, Management
- Differences, Wholesale, Manufact, Fin. Ins.

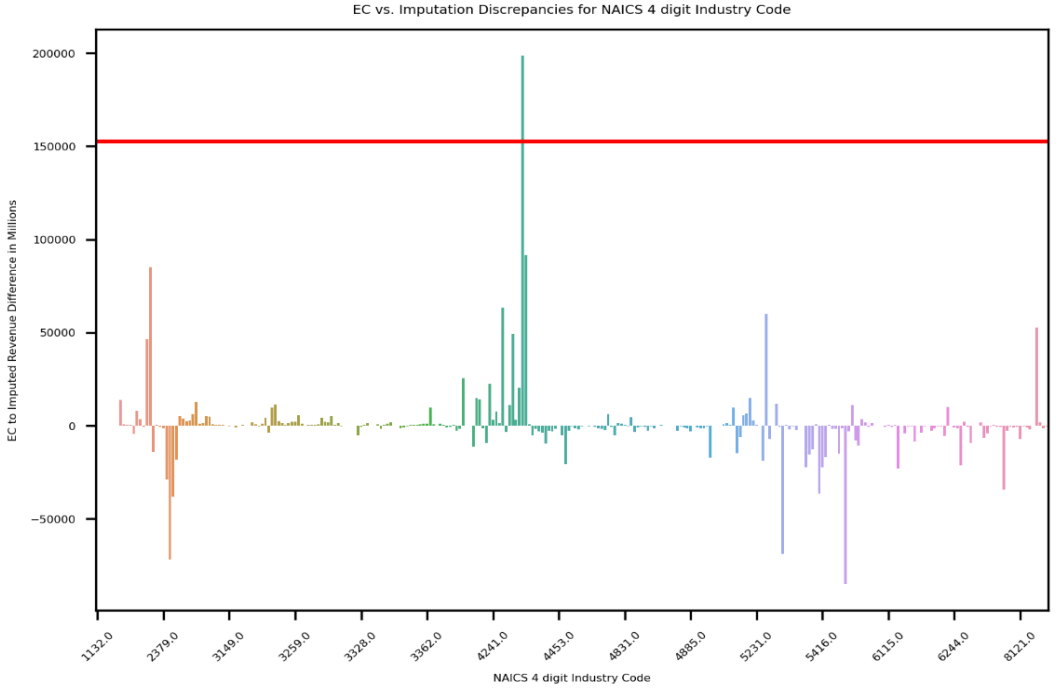




INDUSTRY-SU

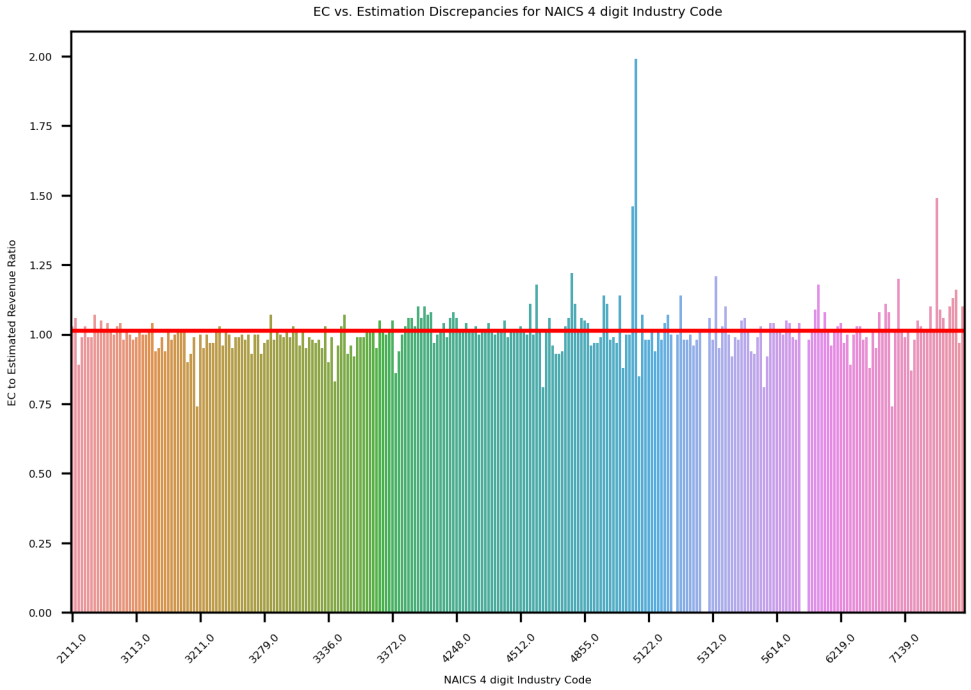
- ❑ Ratios much more variation->next steps
- ❑ Differences, fewer industries are drivers
- ❑ 106/289 industries >20% of EC diff
  - ❑ Sector cancellation

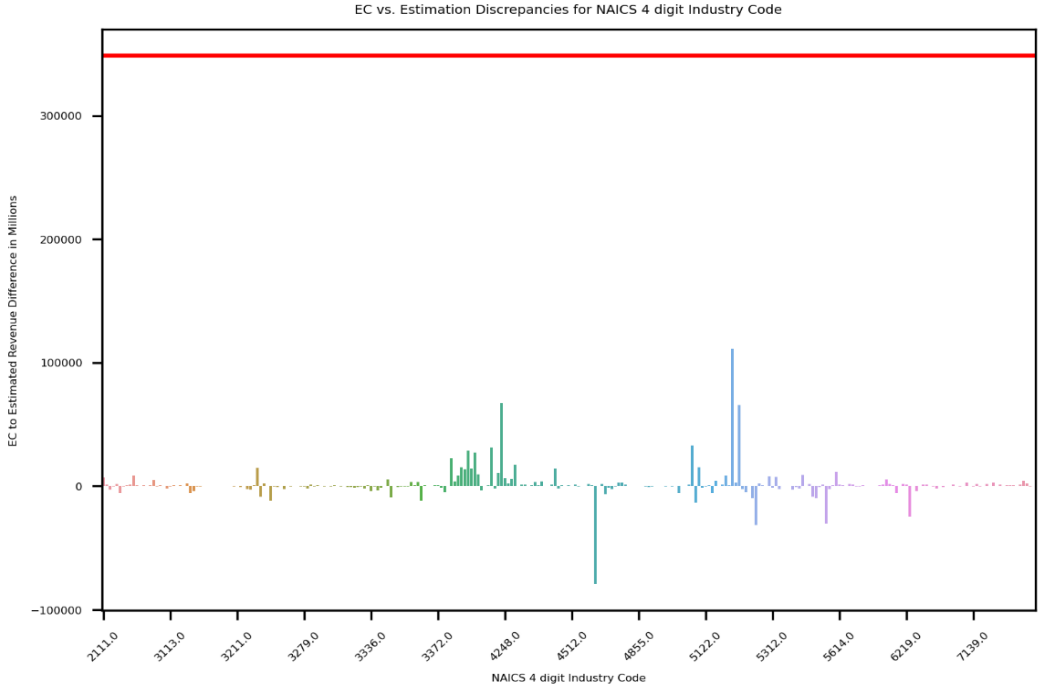




INDUSTRY-MU

- ❑ Post-Raking and Normalization perform very well
- ❑ Still few problematic industries



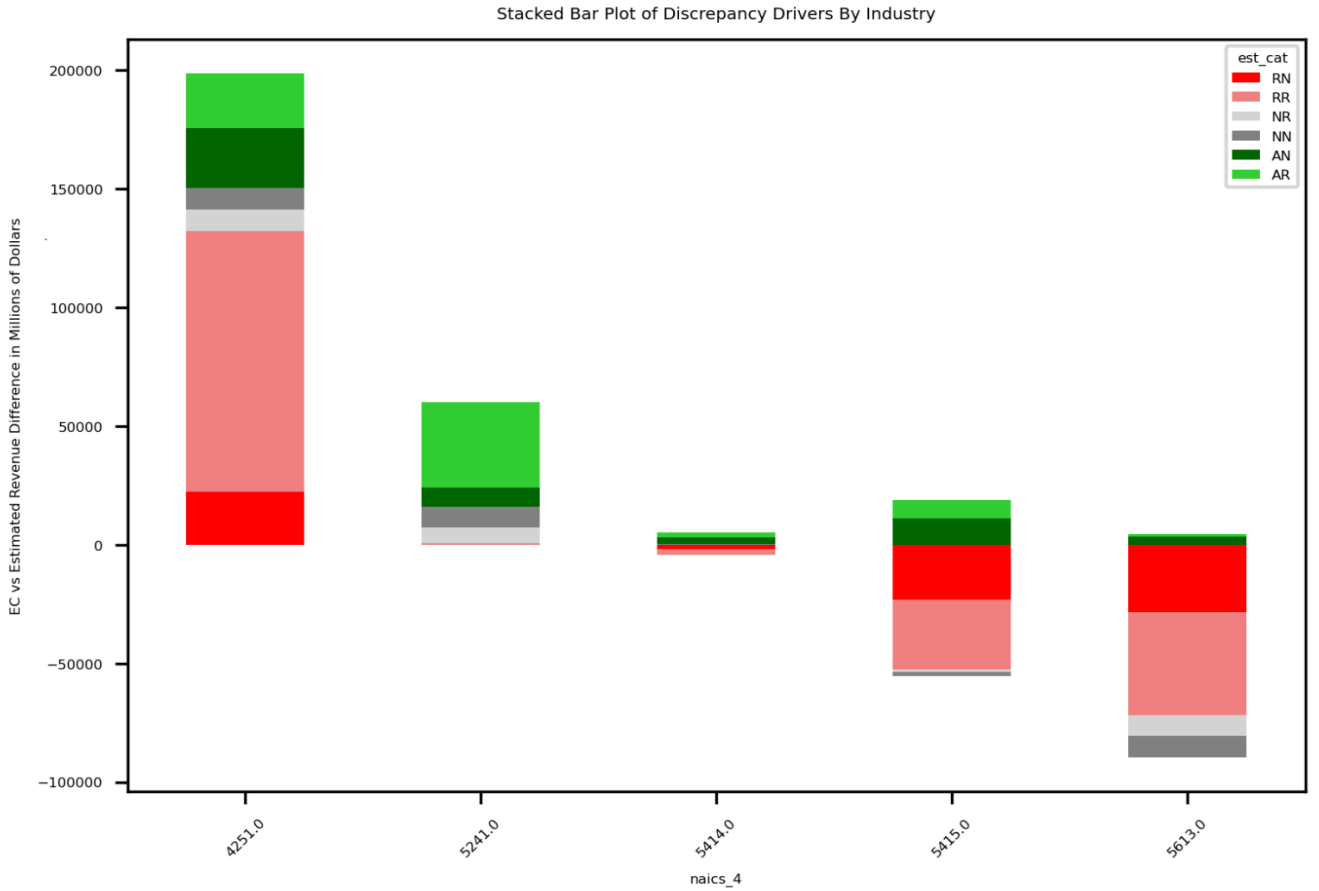


Discrepancy Drivers

- **Problematic Sectors**
  - Wholesale Trade (Biggest Issue)
  - Prof. Sci. Tech
  - Manufacturing
- **Few Industries as Drivers for Sectors**
  - i.e. 4251->Wholesale Electronic Markets and Agents and Brokers
- **Next Steps**
  - EC Imputations vs Our Own
  - Definitional Issues (4251 EC vs. Admin)

ADMIN-SU

4 Digit NAICS Code	Industry
4251	Wholesale Electronic Markets and Agents and Brokers
5241	Insurance Carriers
5414	Specialized Design Services
5415	Computer Systems Design and Related Services
5613	Employment Services



GOALS AND NEXT STEPS

Next Steps

- Refining imputation models based on discrepancy analysis
  - Understanding of imputations within Admin and Census data
- Final deliverable: prototype of establishment-level revenue data series
  - Breaks down national revenue accurately to state and county levels