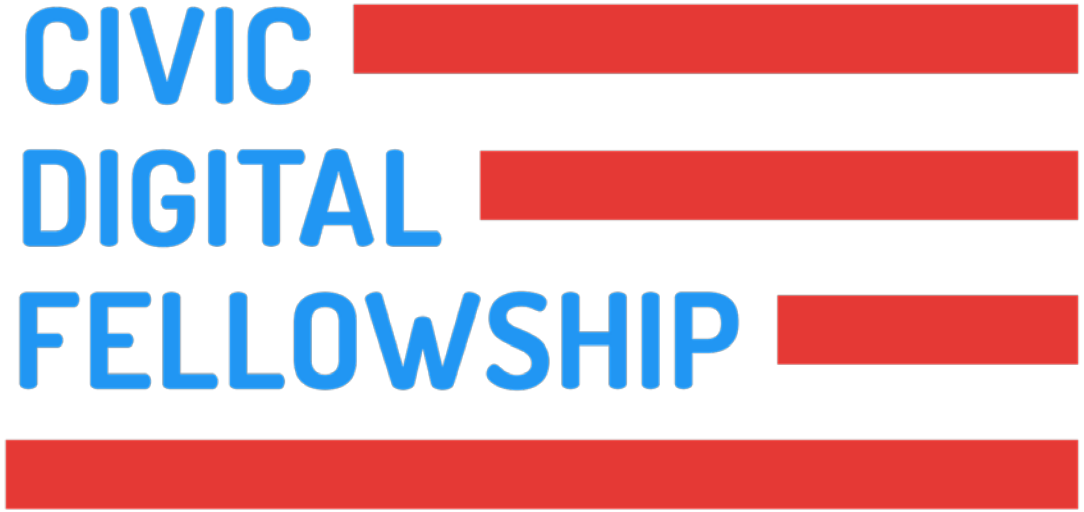


Revenue-Based County Business Patterns

Andy Green and DJ Jain

Supervised by Javier Miranda and Anne Russell

Economy-Wide Statistics Division

Overview / Agenda

- ❑ Motivation and Challenges for the Project
- ❑ Approach and Data Results
- ❑ Conclusion and Next Steps

Motivation and Challenges for the Project

Motivation

The Census Bureau currently creates few products that include revenue data. Revenue is a key economic indicator, and expanding the CBP enables stakeholders to better understand the national economy.

- Census Bureau Uses
  - Expand CBP to deliver a high-quality product, made publicly available
  - Internal use for sampling purposes in annuals
- Bureau of Economic Analysis
  - Characterize enterprises in Small Business Accounts based on annual revenue, instead of # of employees
  - Improves calculations to produce GDP at state and county levels
- Federal Reserve
  - Industrial Production Statistics – relies on CBP to consider industry effects of nationwide events
  - Inform policy using establishment and firm-level revenue changes
- Private Businesses
  - Make planning and operational decisions based on released data of economic activity across industries and geography

Challenges

- Missing data (admin revenue)
  - SU: 1.1M / 5.8M establishments (20%)
  - MU: 53K / 191K firms (27%)
- Unreliable data / outliers
- Structural differences between revenue measures across data sources for certain industries
- Managing data recorded at different levels (establishment vs. tax-entity)

Approach and Results

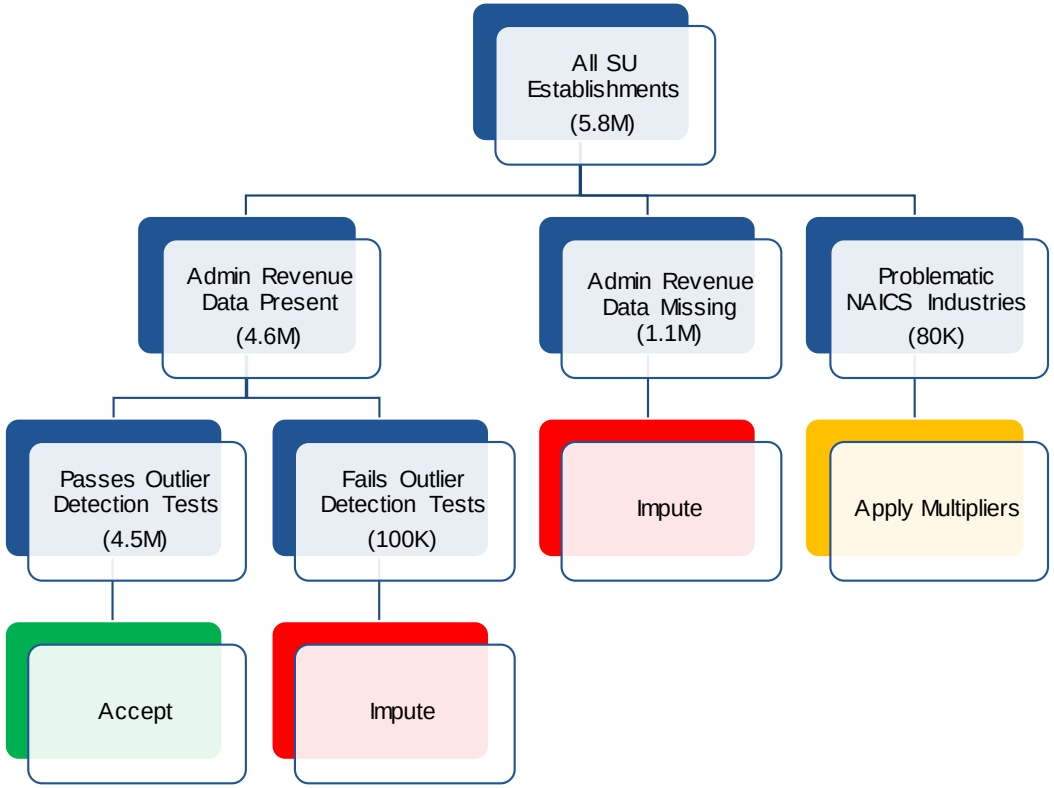
Approach

- Goal: create a revenue CBP file for 2017, using only administrative data for revenue purposes.
- Baseline: CBP microdata for 2017
  - This will provide the full list of relevant establishments, as well as payroll data for imputation purposes
- Revenue data
  - Administrative Data: revenue data and revenue quality flag information will be integrated
  - Economic Census Data: revenue data will solely be used for comparison/verification purposes with the administrative data; no imputations will come from these data
- Revenue verification and imputation will be handled separately for SU and MU establishments

Establishment Type	# of Establishments in CBP Microdata
SU	5.8M
MU	2.1M
Total	**7.9M**

Approach – SU Establishments

- Outlier detection
  - Revenue quality flags
  - Additional outlier detection rules
- Imputation algorithms
  - ASE-based approach
  - Regression model
- Industry deep-dive comparison of Economic Census revenue vs. Administrative revenue

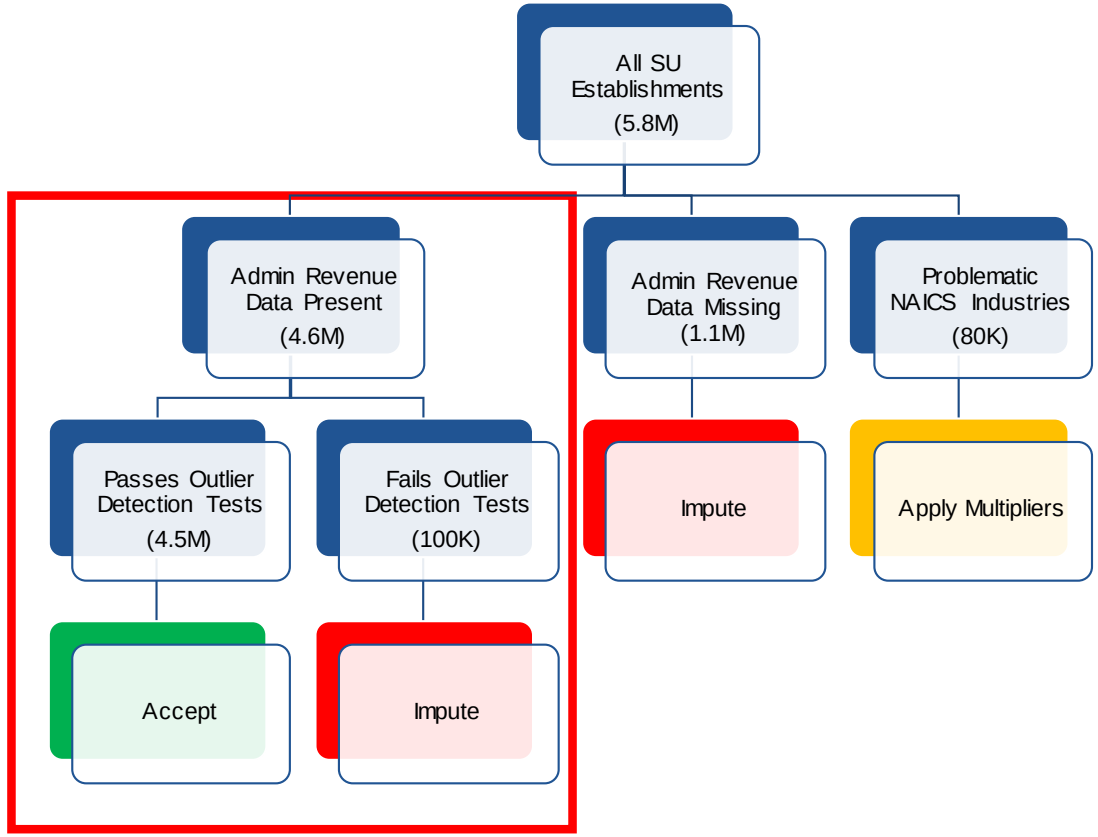


SU Establishments – Outlier Detection

Challenge

- ❑ Analyzing discrepancies between Administrative data revenue and Economic Census revenue → high average discrepancies
- ❑ Driven by a relatively small number of massive discrepancies
- ❑ Need to find some way to identify these outliers, without explicitly filtering by the discrepancy

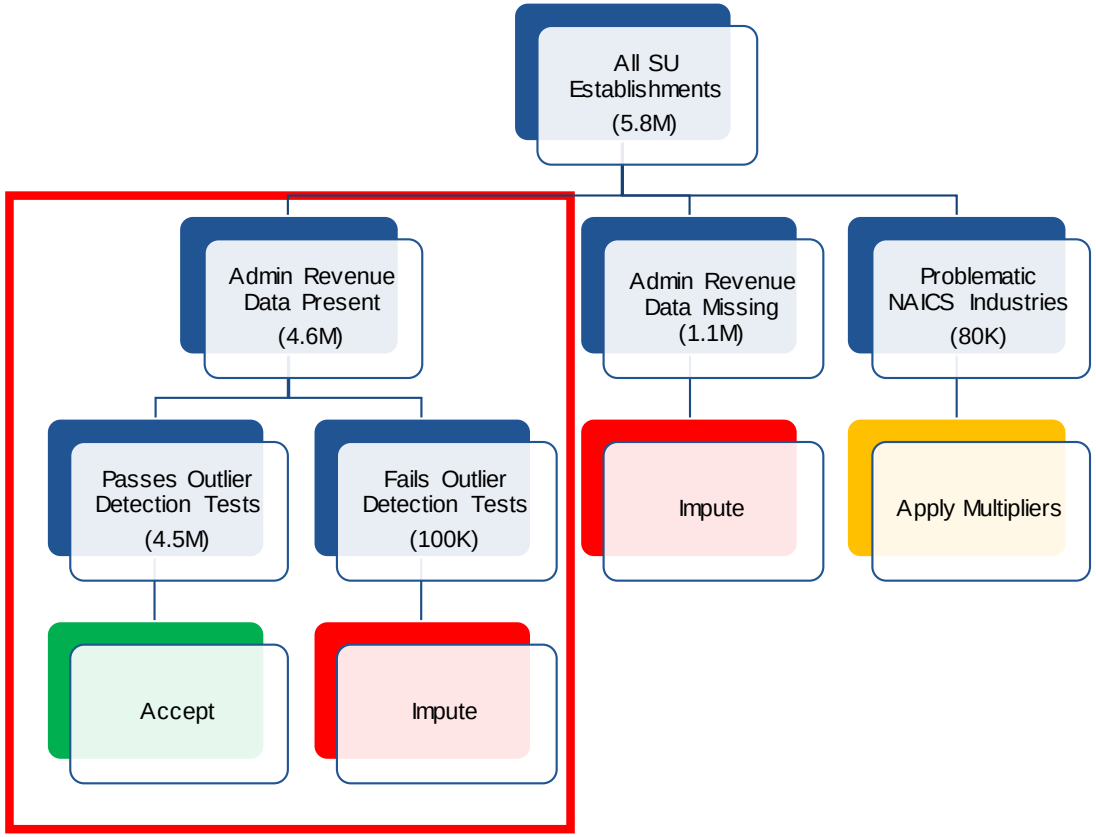
Metric	Result
Total Revenue % Difference (Admin vs. EC)	40% higher
Average Establishment-Level $ Difference (Admin vs. EC)	$681K higher



SU Establishments – Outlier Detection

Approach

- Leveraging "revenue quality flags" created by the BR team (ESMD)
  - Revenue quality flags of A/S/T produce much smaller discrepancies than U flags
- Implementing additional outlier detection rules:
  - Revenue-payroll ratio outliers
  - Payroll-employment ratio outliers
  - Suspiciously large births and suspiciously large growth rates

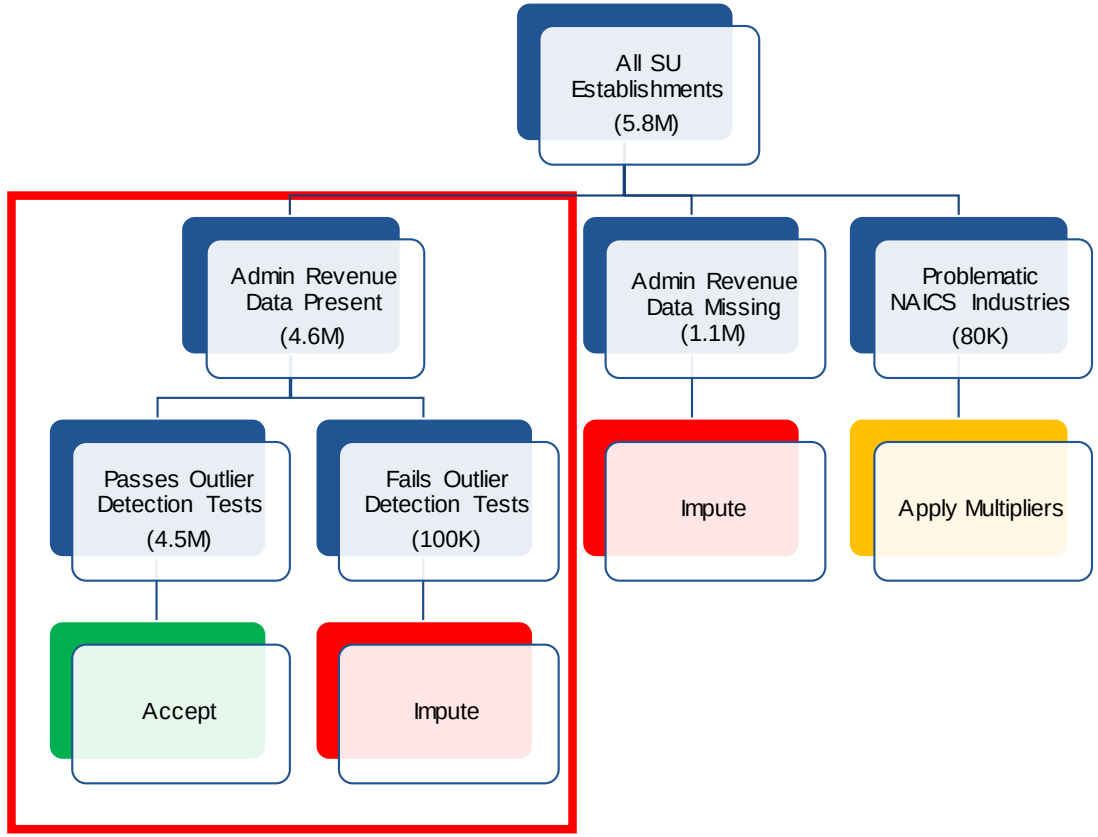


SU Establishments – Outlier Detection

Results

- ❑ Establishments that passed all outlier detection tests
- ❑ Establishments that failed any of the outlier detection tests

	Pass	Fail
Total Revenue % Difference (Admin vs. EC)	1.2% higher	220% higher
Average Establishment-Level $ Difference (Admin vs. EC)	$16K higher	$27.3M higher



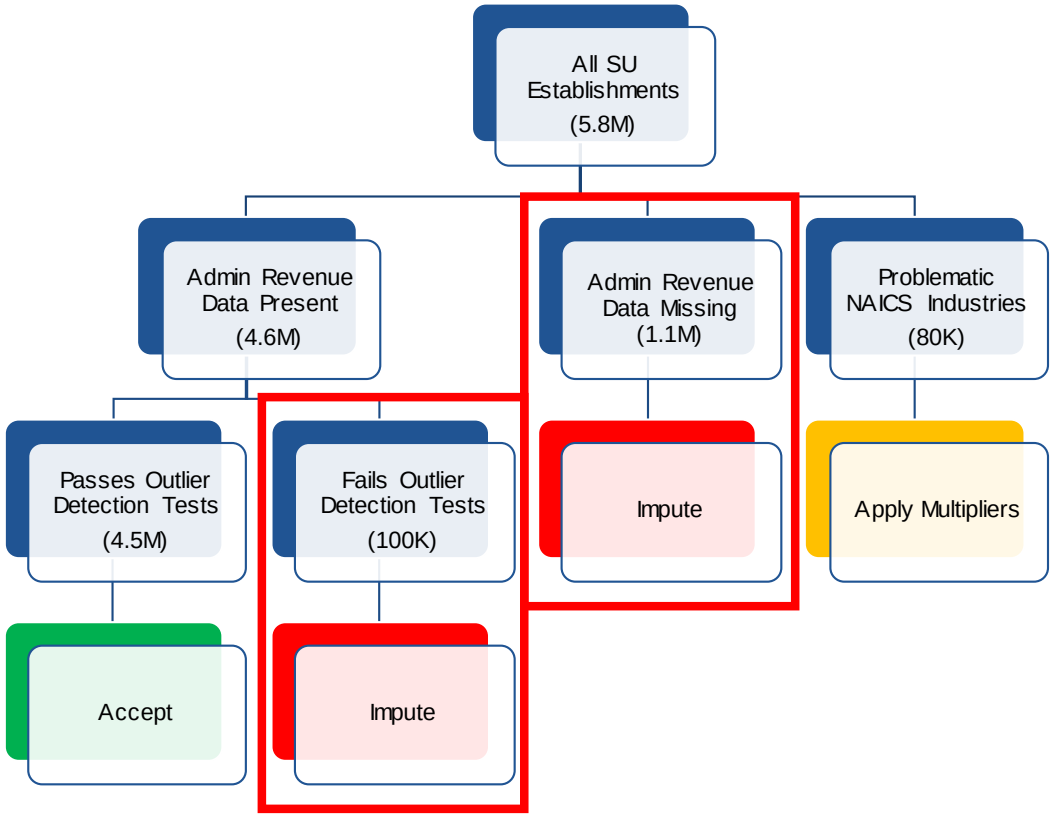
SU Establishments – Imputation

Challenge

- ❑ Outliers (100K)
- ❑ Missing data (1.1M)

Approach

- ❑ ASE-based approach
  - ❑ Revenue-payroll ratios, calculated by:
    - ❑ Most complex: 4-digit NAICS, state, firm size
    - ❑ Medium complex: 4-digit NAICS, state
    - ❑ Least complex: 4-digit NAICS
  - ❑ These ratios are applied to actual payroll from CBP to get projected revenue
- ❑ Regression model

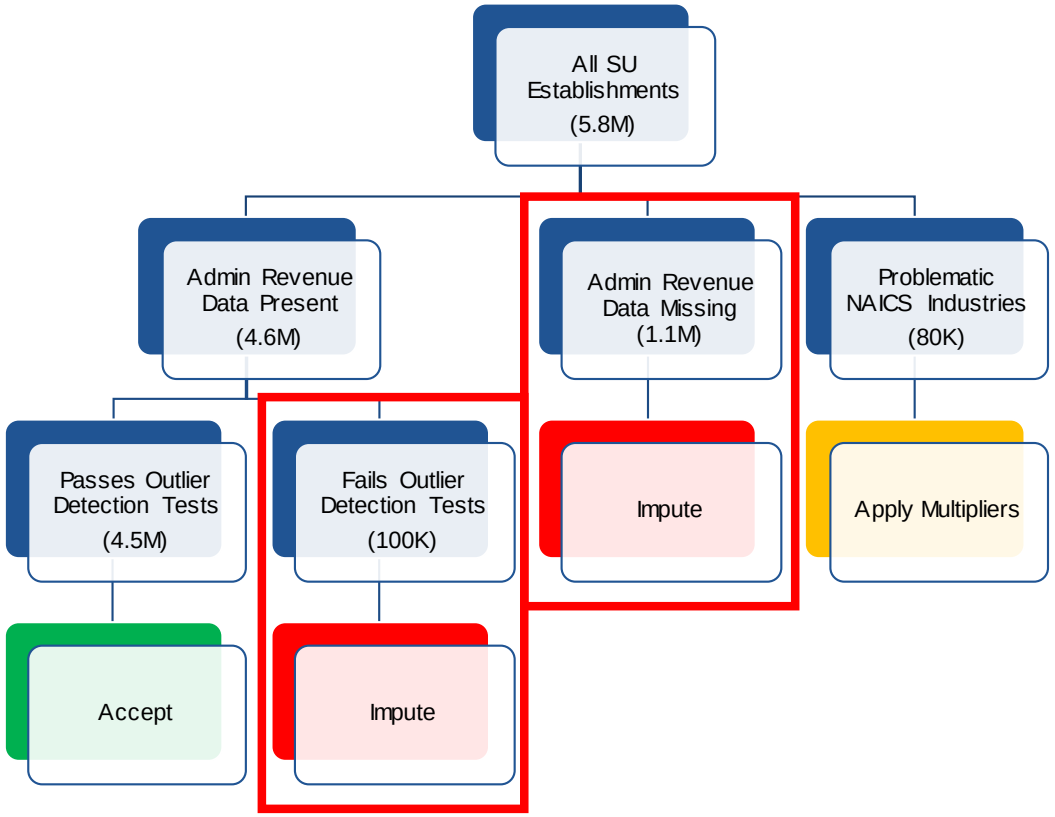


SU Establishments – Imputation

Results

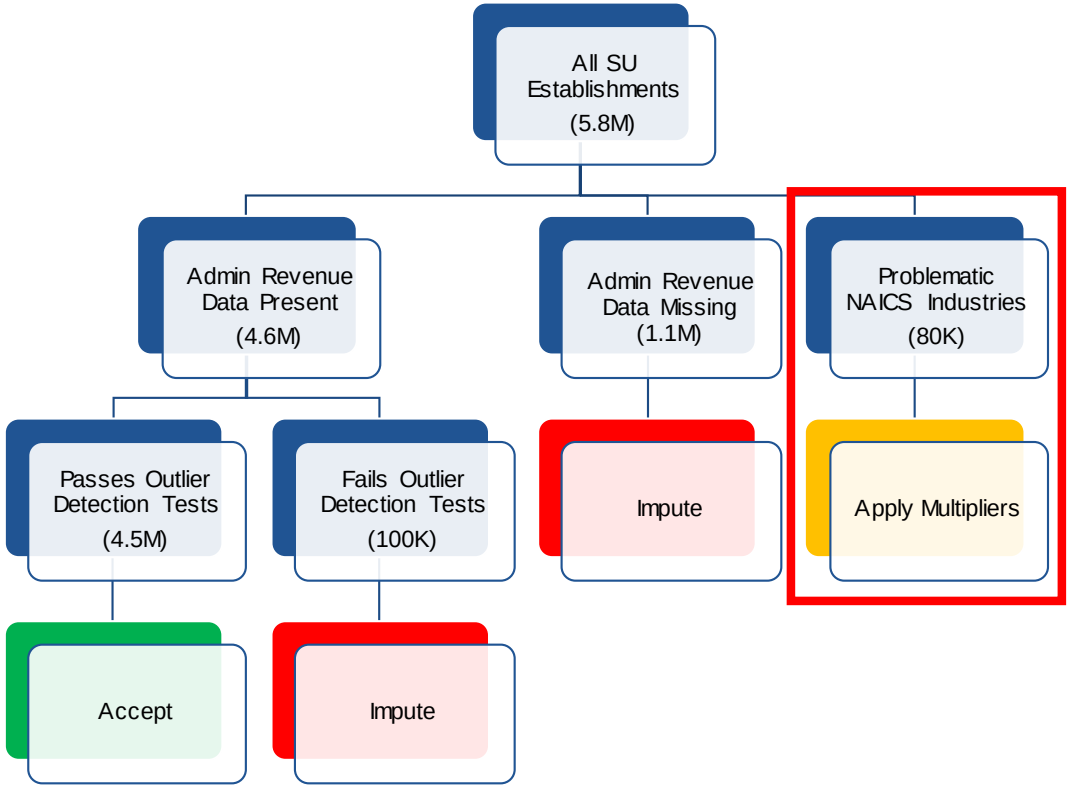
Outliers	Before	After
Total Revenue % Difference (Admin vs. EC)	220% higher	12% lower
Average Establishment-Level $ Difference (Admin vs. EC)	$27.3M higher	$1.5M lower

Missing Data	Before	After
Total Revenue % Difference (Admin vs. EC)	-	3.6% lower
Average Establishment-Level $ Difference (Admin vs. EC)	-	$44K lower



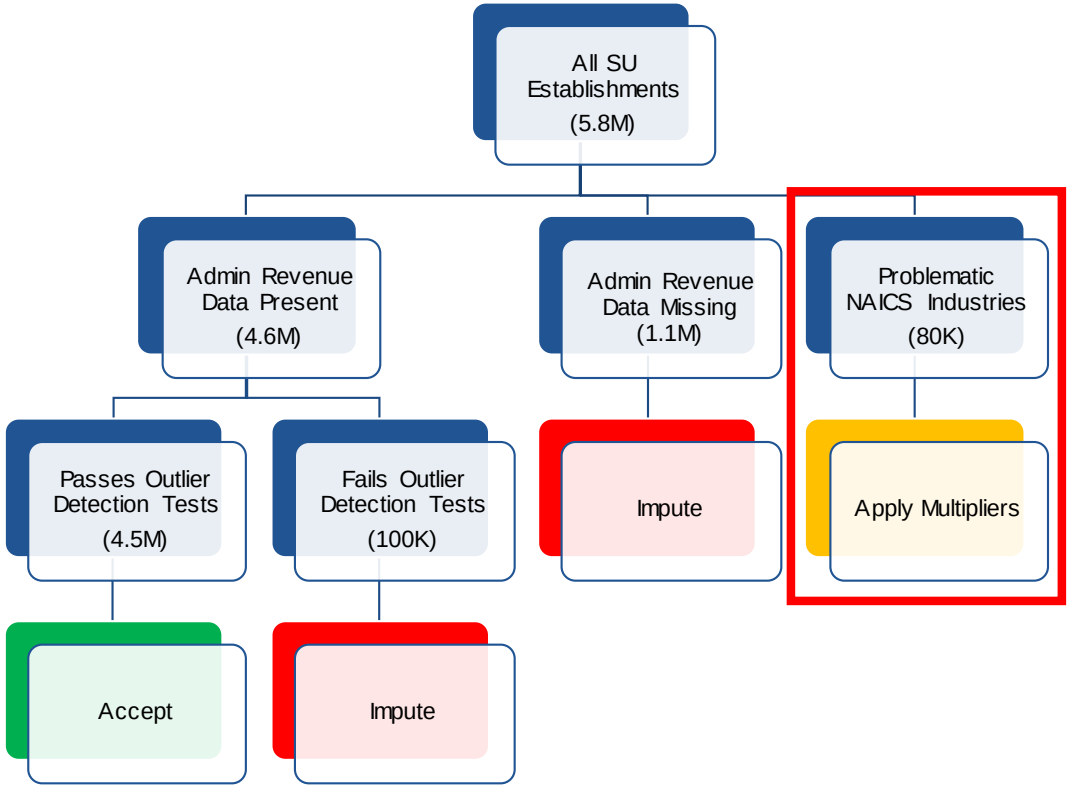
Industry Adjustment: Goals

- Pinpoint structural & systematic differences between Respondent Economic Census (2017) revenue and Admin Receipts
- Investigate discrepancies (ranked on average by establishment) for 4-digit NAICS industries both in problematic 2-digit sectors as well as combined dataset as a whole
  - Explore whether specific 4-digit industries are the root of large discrepancies within otherwise unproblematic 2-digit sectors
- Develop adjustment ratios to calibrate Admin Receipts with Respondent EC, to produce final revenue value for CBP



Industry Adjustment: Approach (SUs)

- Outlier removal: using payroll-revenue ratios and removing tails based on means and standard deviations from each 4-digit industry
- Preliminary investigation: SUs in sectors 42 and 48 (as flagged by Revenue Comparison Report document)
  - Plot distribution, compare mean and dispersion, flag outliers in either dataset (in conjunction with analyzing 4-digit industries within these sectors)
- Repeat steps for problematic industries with an unproblematic 2-digit sector (ex: retail) to compare distribution shapes between EC and Admin
- Analyze aggregate revenue between EC and Admin at every 4-digit industry level, develop averages of discrepancy based on number of establishments



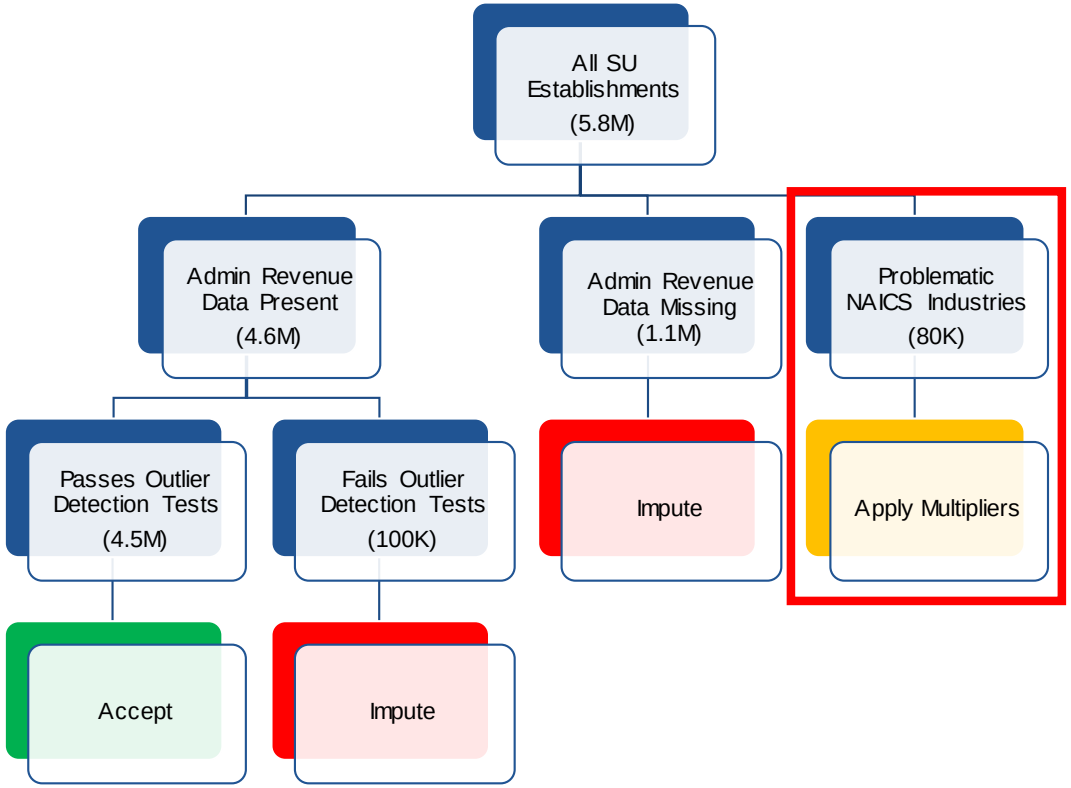
Industry Adjustment: Results (SUs)

				DIFF_DOLLAR			ESTAB_COUNT		CENSUS_ADMIN_RATIO		
NAICS_4DIGIT	NAICSECN	RCPTOT	BESTADMIN_RCPT_2017	ABS	MEDIAN	AVERAGE	FILTERED	RAW	MED	AVG	AVG_W
4251	42	$53.1 bil	$10.00 bil	$43.1 bil	$575	$2.9k	14.7k	16.1k	2.94	8.65	5.30
4885	48	$6.36 bil	$11.80 bil	$5.44 bil	$125	$1k	5.0k	5.3k	0.90	0.77	0.54
5418	54	$11.68 bil	$13.29 bil	$1.61 bil	$20	250	6.3k	6.6k	1.00	1.02	0.88

- 4251: Wholesale Electronic Markets and Agents and Brokers
- 4885: Freight Transportation Arrangement
- 5418: Advertising, Public Relations, and Related Services

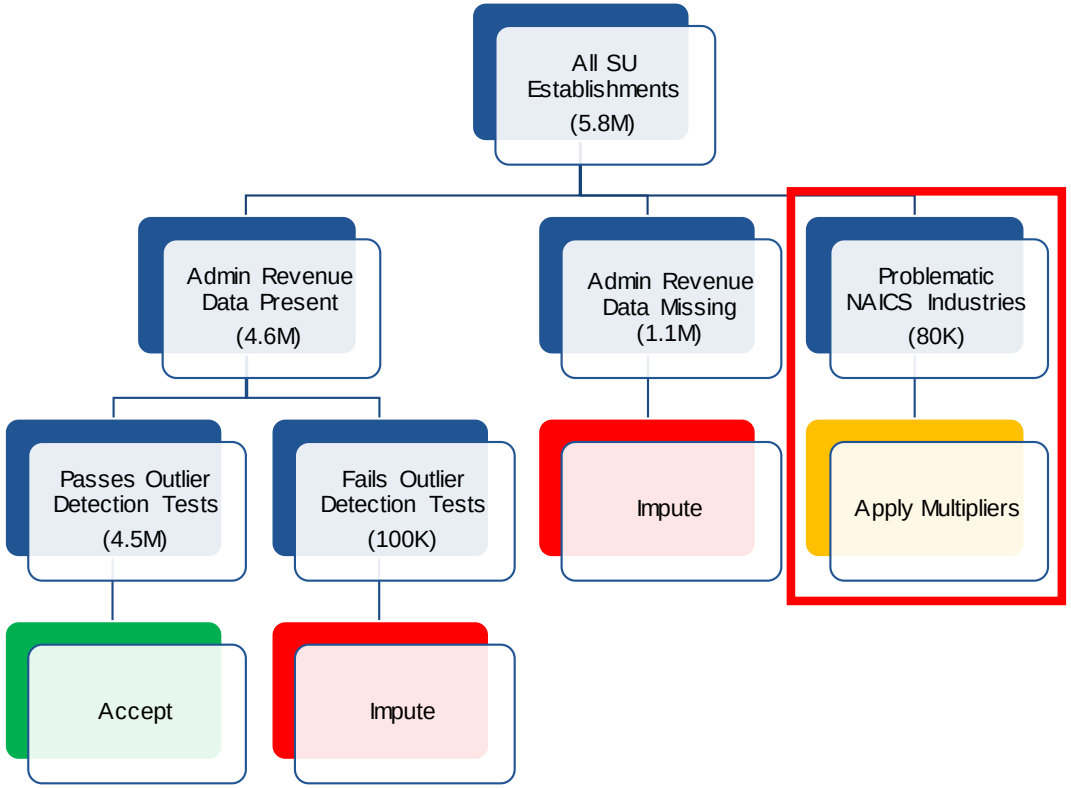
Industry Adjustment: Approach (MUs)

- Limit to firms with all EINs having a revenue amount
  - Completely exclude MUs with EINs that have missing revenue (inlier issue)
- Only consider firms where all establishments under a single EIN are in the same 4-digit industry
  - Treat EINs with different industries as own pseudo-firms for aggregate comparison purposes
  - Do not aggregate to firm level for different industry EINs



Industry Adjustment: Challenges

- Developing business rules to select group of industries to adjust revenue
  - Advised by trends in “control” unproblematic industries
- Small cells with few establishments and high average discrepancies
  - Set aside industries, do not greatly affect final aggregate revenue calculations in CBP product
- MUs: firms may be comprised of EINs with different 4-digit industries, revenue aggregates differ based on Census response and financial engineering
  - Using MUs to corroborate findings in SU problematic industries
- Adjustment ratios could vary based on establishment size, legal form

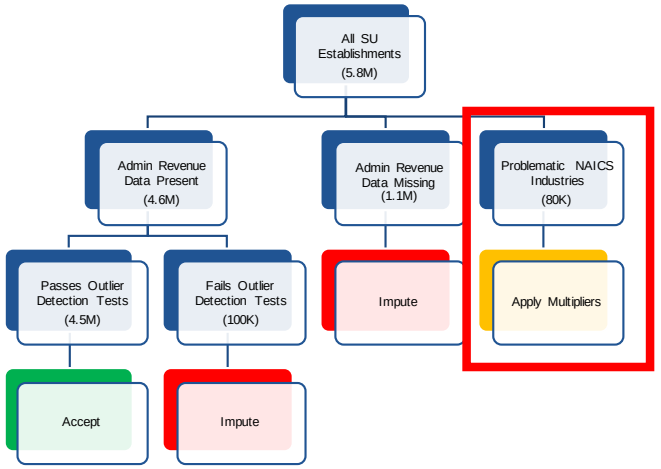


SU Establishments – Industry Adjustments

Results

- ❑ Problematic NAICS
- ❑ Admin Revenue Present
- ❑ Passes Outlier Detection

	Before	After
Total Revenue % Difference (Admin vs. EC)	49% lower	0%
Average Establishment-Level $ Difference (Admin vs. EC)	$1.4M lower	$0



- ❑ Problematic NAICS
- ❑ Admin Revenue Present
- ❑ Fails Outlier Detection

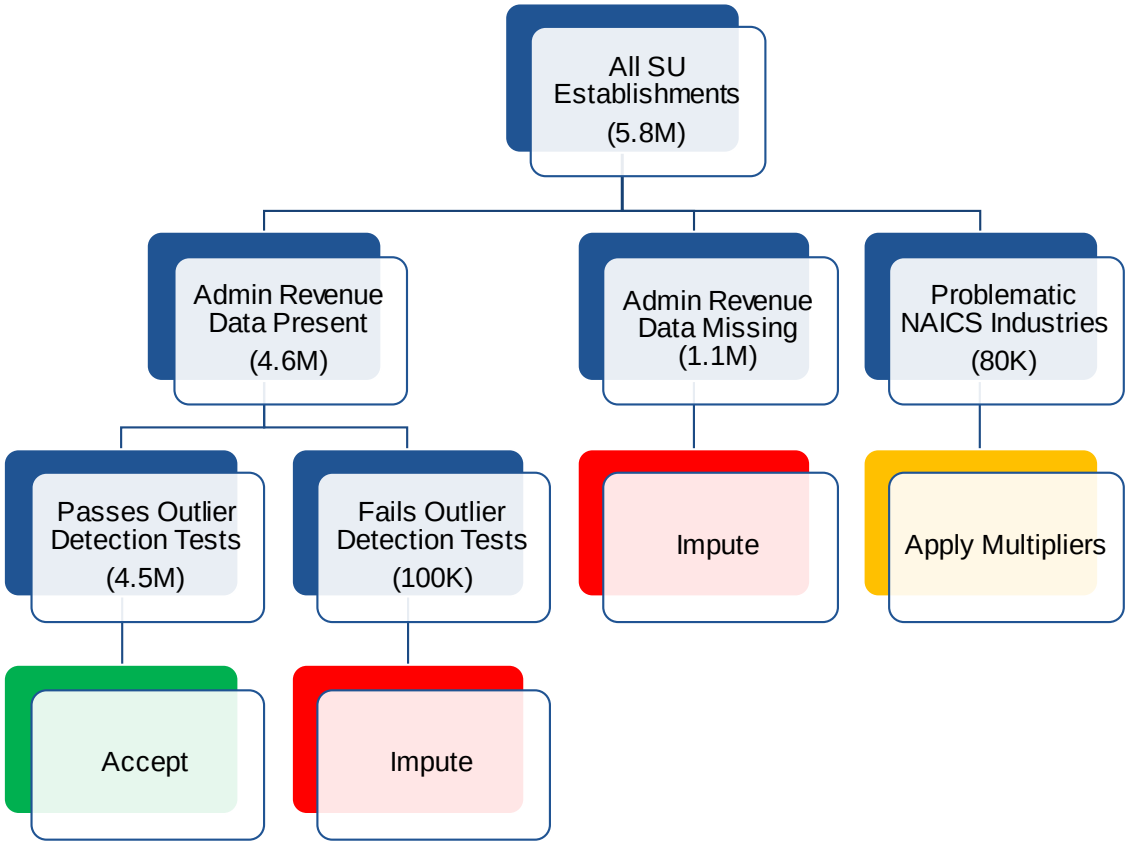
	Before	After
Total Revenue % Difference (Admin vs. EC)	13% higher	6% lower
Average Establishment-Level $ Difference (Admin vs. EC)	$4.6M higher	$2M lower

- ❑ Problematic NAICS
- ❑ Admin Revenue Missing

	Before	After
Total Revenue % Difference (Admin vs. EC)	-	22% higher
Average Establishment-Level $ Difference (Admin vs. EC)	-	$866K higher

Conclusion and Next Steps

SU Establishments – Final 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

Next Steps + Improvements

Multi-Unit Establishments

- Disaggregate revenue from firm to establishment level
- Refine revenue for large firms with unlinked variables
- Create aggregate firm value tabulations

Overall Improvements

- Implement more advanced imputation approaches: regression, CART models
- Special processing/refinement for certain industries and geographies

Industry Adjustment

- Create MU adjustment ratios based on existing framework
- Generate specific adjustment ratios based on establishment size

Thank you!

Appendix

SU Establishments – Outlier Detection

- Approach
  - Leveraging "revenue quality flags" created by the BR team
  - Revenue quality flags of A/S/T produce much smaller discrepancies than U flags
  - However, there are still clearly a relatively small number of observations doing a significant amount of damage to the means → need to develop additional outlier detection rules

	Total	Acceptable	Small	Tabulate	Unacceptable
count	4.6M	3.3M	615K	562K	112K
mean	$938	$94	$21	$500	$33K
25%	$0	$0	$0	$0	$31
50%	$0	$0	$0	$0	$634
75%	$6	$0	$1	$136	$2,881

SU Establishments – Outlier Detection

- Approach
  - Need to develop additional outlier detection rules:
    - Revenue-payroll ratio outliers
    - Payroll-employment ratio outliers
    - Suspiciously large births and suspiciously large growth rates
  - Treat these outliers as missing data

	Total	Acceptable	Small	Tabulate	Unacceptable
count	4.6M	3.3M	613K	559K	101K
mean	$260	$93	$21	$483	$5,702
25%	$0	$0	$0	$0	$8
50%	$0	$0	$0	$0	$537
75%	$6	$0	$1	$135	$2,486

SU Establishments – Imputation

Results – Comparing Approaches:

ABS	Best Available	Most Complex	Medium Complex	Least Complex
count	3.29M	3.28M	3.29M	3.29M
mean	$392	$387	$410	$420
25%	$68	$67	$68	$69
50%	$160	$160	$163	$165
75%	$365	$363	$393	$400
RAW	Best Available	Most Complex	Medium Complex	Least Complex
count	3.29M	3.28M	3.29M	3.29M
mean	$1	$0	$0	$0
25%	-$115	-$115	-$146	-$147
50%	$36	$36	-$3	-$2
75%	$193	$193	$182	$187

ABS	Regression
count	3.3M
mean	$599
25%	$116
50%	$260
75%	$542
RAW	Regression
count	3.3M
mean	$0
25%	-$223
50%	$17
75%	$306

Background Info on the Economic Census, BR, and CBP

Economic Census

The Economic Census, conducted every five years, surveys U.S. businesses to quantify economic activity across industries and regions.

- Functions: compiles information regarding approximately 4 million businesses, updates CBP and BR with births/deaths, informs national GDP calculations
- Responses include: EIN, physical location, primary business activity (based on employment), sales/revenue/receipts, employment and payroll
- External uses: impacts business decisions regarding operations, investments, and development

Reference: About the Economic Census

Business Register

The BR is a comprehensive multi-relational database that contains a record for each known legal establishment. It serves as the "phone book" of all business establishments in the US.

- Functions: data frames for economic surveys and censuses, central repository for admin. records, source data for CBP, ZIP Business Patterns, and BDS
- Statistical unit: establishment (primary), EIN, enterprise
- Variables: geography, industry/organizational, measures of business activity
- Sources: IRS, BLS, SSA, Census Bureau

County Business Patterns

County Business Patterns (CBP) is an annual series that provides subnational economic data by industry.

- Metrics: # of establishments, employment (as of March 12), first quarter payroll, annual payroll
- Granularity:
  - Industry: 6-digit NAICS industry code
  - Geography: national, state, county, metropolitan area, zip code, congressional district
- Used by: federal agencies, state and local governments, businesses, researchers/academics
- Sources: BR, Economic Census, annual surveys

Reference: County Business Patterns