Differential Privacy, Differential Accuracy?: A Use Case on Data Segmentation by Tract-level Tenure Majority

Leslie Reynolds and Jan Vink
Program on Applied Demographics
Cornell University

Introduction

- People in areas with mostly rental units have different characteristics than those in areas with mostly owner-occupied housing units
    - age of householder, income level, family size etc.
- Research on differentials by tenure is essential for many government/social programs
    - e.g. HUD rental assistance program requirements are determined by research on tenure by household characteristics
- Census demonstration data (March 2022) of the total population appear accurate to the original file
    - *But:*
        - Areas with mostly renter-occupied and mostly owner-occupied units are geographically scattered
        - Don’t necessarily fit within the top-down approach of differential privacy

Research Questions

1. Does the top down algorithm impact the comparisons between aggregate groups?

2. Does the top down algorithm differentially impact accuracy within tenure groups?

3. Does geographic scale matter?

Data and Methods

- 2010 Summary File 1 and 2010 Differential Privacy Demonstration data file (released March 29, 2022), retrieved from IPUMS[1]

- Census tract-level Housing Unit and Person files, merged by geocode
  - Excluded Puerto Rico
  - Only kept tracts with at least 200 households (to exclude special purpose tracts)

➢ Result: 71,842 tracts in the U.S., 4,772 tracts in New York State, 189 in Monroe County (Rochester), and 139 in Onondaga County (Syracuse)

1. David Van Riper, Tracy Kugler, and Jonathan Schroeder. IPUMS NHGIS Privacy-Protected 2010 Census Demonstration Data, version 20220310 [Database]. Minneapolis, MN: IPUMS. 2020.

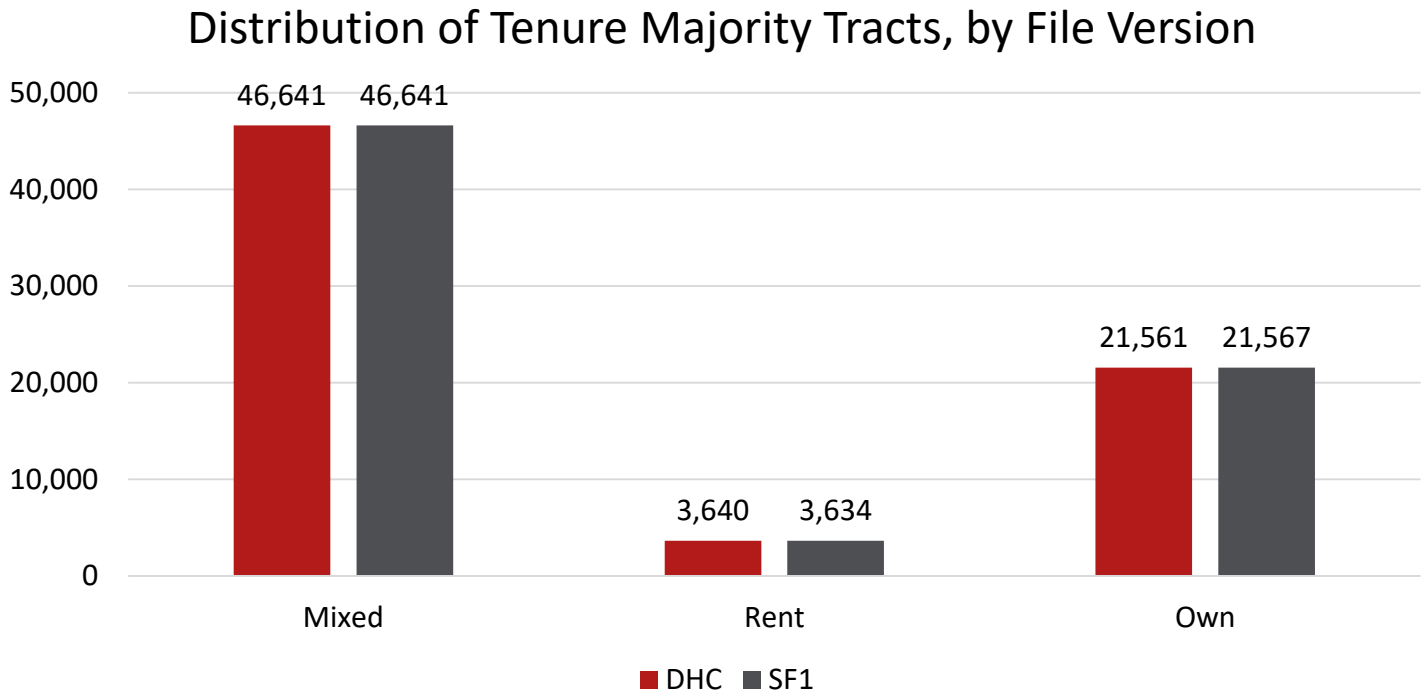
Data and Methods (cont.)

- Household variables: Non-family households, single person households, large households (5+ residents), and households with children (under 18)
- Population variables: Children under 5, children 5-14, adults aged 18-24 and 65-74, women aged 18-24 and 65-74, and men aged 18-24 and 65-74

- Analytical Variable: dominant tenure category in a Census tract
    - Majority owned (>=80% owned households)
    - Majority rented (<=20% owned households)
    - Mixed tenure (> 20% & < 79% owned households)
- Percent owned Calculated as [(owned + owned with mortgage)/total occupied households] *100

Data and Methods (cont.)

- Original SF1 and Demonstration DHC had almost identical distributions of tracts across tenure majority categories



Selected Metrics

- Average count and percent difference from the original data values
  - Mean Error: $\frac{1}{N}\sum(X_{dp} - X_{sf})$
  - Mean Algebraic Percent Error (MALPE): $\frac{1}{N}\sum\frac{X_{dp}-X_{sf}}{X_{sf}}$ *100%
- Degree of error between the original and demonstration data
  - Median Absolute Percent Error (MdAPE): Median $(|\frac{X_{dp}-X_{sf}}{X_{sf}}|)$*100%
- Precision of the estimates
  - “Big” Errors: Mean Absolute Error >=10 & Mean Absolute Percent Error >=10%

Broad Scope Comparison of the Files

- Large-scale descriptive statistics showed few differences between files
- Proportions of households with selected characteristics were nearly identical between files
  - Some variation across tenure-majority types and analytic areas
  - Example: Figures 2 and 3

Figure 2: Shares of Households with Children

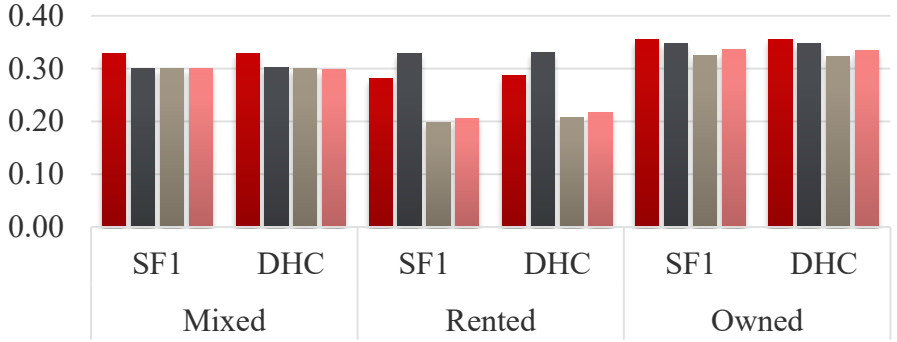
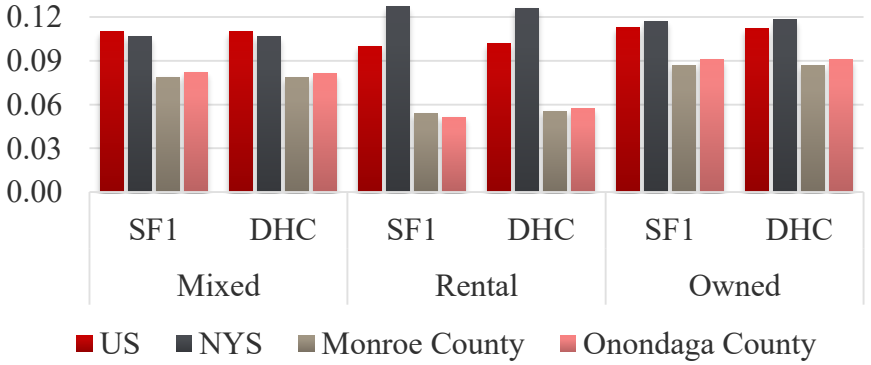


Figure 3: Shares of Households with 5+ Residents



Broad Scope Comparison(cont.)

- Similarities between file versions also observed for the Person files
  - Differences across tenure and area, but patterns mirrored between files

Figure 4: Shares of the Population Aged 5-14

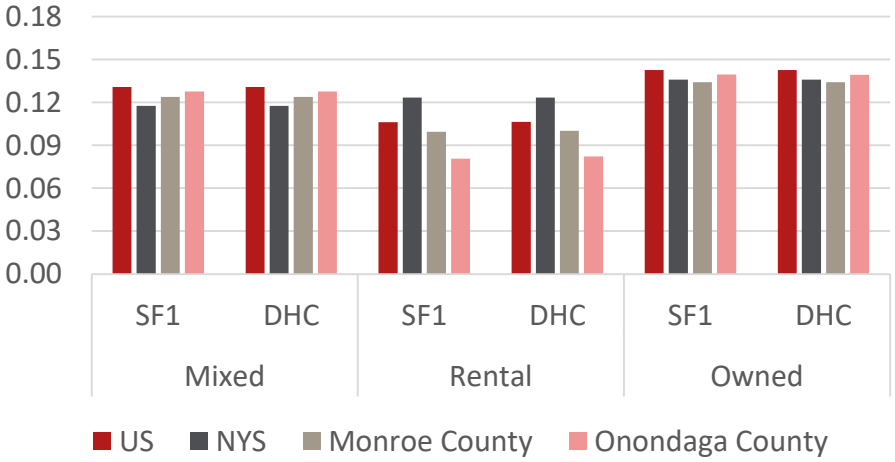
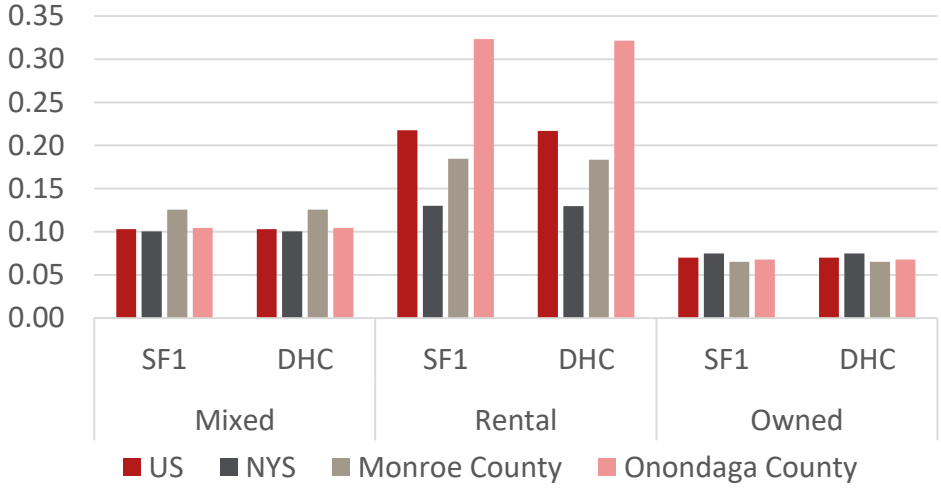


Figure 5: Shares of the Population Aged 18-24



Broad Scope Comparison of the Files (cont.)

- Without separation by tenure, count discrepancies are slight
  - Non-family households: largest mean error but small MALPE
    - On average, 1.53 more non-family households were counted in SF1 than in the DHC.
  - Large households had the highest MALPE; 5.0%

Table 1: Differences in the Mean Counts, Mean Error, and Mean Percent Error Between 2010 SF1 and Demonstration DHC

	Mean			
Measures:	DHC	SF1	Error	% Error (MALPE)
Households with children	486.23	485.58	0.65	1.14
Large Households	172.84	172.90	-0.06	5.00
Single Person Households	446.26	446.23	0.03	0.34
Non-Family Households	557.22	558.75	-1.53	0.50

Results: Bias in the Household File

Table 2: Mean Error by Tenure Majority and Geographic Area

	Geography	Mixed	Rented	Owned
Nonfamily Households	United States	**-2.02***	**-6.10***	**-0.33***
	New York State	**-1.44***	**-2.03***	**-1.34***
	Monroe County	-1.98	-6.92	0.58
	Onondaga County	-0.67	-10.79	0.54
Single Person Households	United States	-0.01	**-1.93***	**0.38***
	New York State	0.11	-0.38	0.01
	Monroe County	0.80	-6.08	0.79
	Onondaga County	0.96	**-4.79***	0.65
5+ Person Households	United States	-0.14	**2.56***	-0.01
	New York State	-0.16	**-1.67***	**1.27***
	Monroe County	-0.92	2.00	-0.38
	Onondaga County	-1.17	5.79	-0.33
Households with children	United States	**0.69***	**6.08***	**-1.86***
	New York State	0.69	**2.77***	**-0.73***
	Monroe County	-1.76	10.46	**-3.42***
	Onondaga County	-3.85	11.00	-2.04

*Errors significantly different from zero are bolded and denoted by an asterisk

Results: Accuracy in the Household File

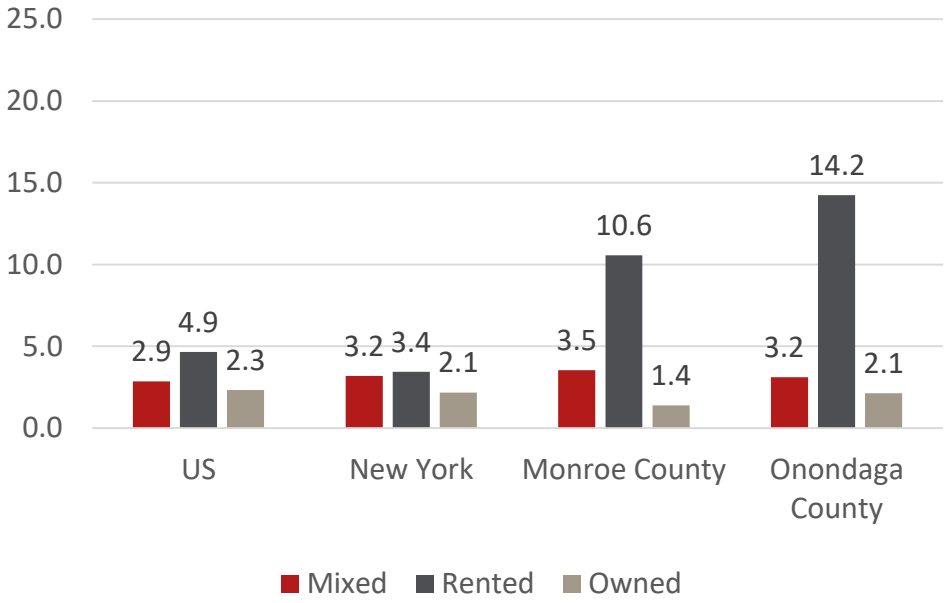
- MdAPE was largest for households with 5+ residents and households with children

Households with children:

- Differences between files were smallest in owner majority areas
  - Minimum: Owner majority areas of Monroe County (1.4%)

- Rental majority areas had most noticeable accuracy issues across all aggregate levels
  - Maximum: Rental majority areas of Onondaga County (14.2%)

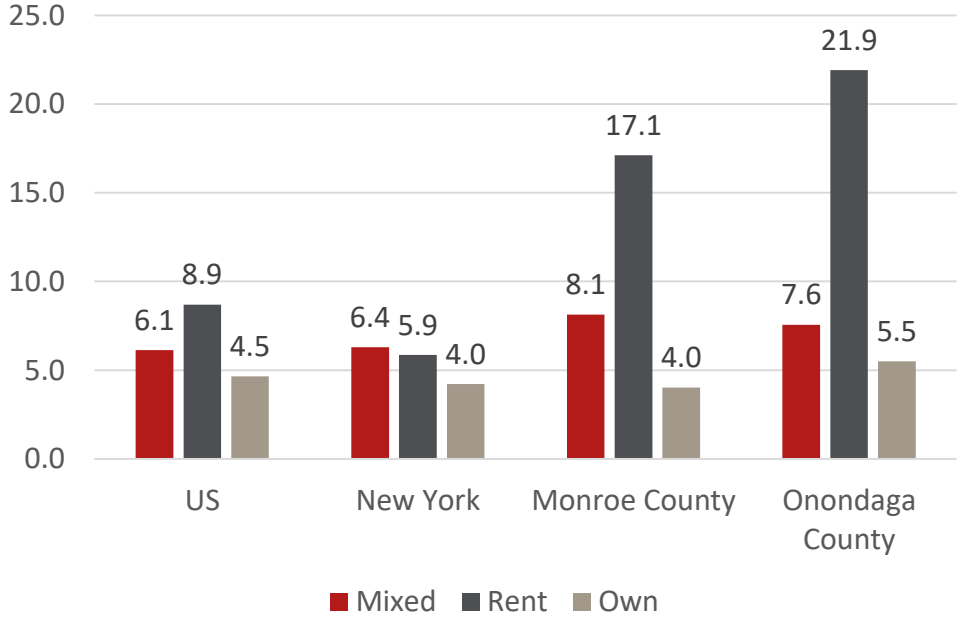
Figure 6: Median Absolute Percent Error (MdAPE), Households with Children



Results: Accuracy, Household File (cont.)

- Large households had even more extreme values
- Owner majority areas had the lowest MdAPE
  - Minimum: 4.0%, owner majority areas of New York State and Monroe county
- Rental areas for each aggregate level had the highest MdAPE
  - Maximum: 21.9%, rental majority areas of Onondaga county
  - Exception: mixed tenure areas of New York state, MdAPE= 6.4%

Figure 7: MdAPE, "Large" Households (5+ people)



Results: Precision in the Household File

- Large Households and Households with children had the largest shares of tracts with big errors
- United States:
  - rental majority areas were least precise, owner majority areas were most precise
- New York State:
  - Mixed tenure areas were least precise, owner majority areas were most precise

Table 3: Share of Tracts in the U.S. with Big Errors

	Single Person HH	Nonfamily HH	5+ Person HH	Households With Children
Mixed	2.6%	3.2%	26.8%	7.2%
Rent	5.2%	7.0%	36.8%	27.4%
Own	2.9%	3.5%	16.8%	4.3%

Table 4: Share of Tracts in New York State with Big Errors

	Single Person HH	Nonfamily HH	5+ Person HH	Households With Children
Mixed	3.9%	6.1%	27.8%	8.7%
Rent	3.4%	3.7%	23.0%	15.0%
Own	2.0%	2.0%	13.6%	2.3%

Results: Precision in the Household File(cont.)

- Onondaga County had the highest shares of tracts with big errors of all aggregate levels
- Rental majority areas were most problematic for:
  - 5+ person households: 46.2% in Monroe, 50% in Onondaga
  - Households with children: 53.8% in Monroe, 64% in Onondaga
- Owner majority areas were generally most precise; more variation for sub-state aggregate levels

Table 5: Share of Tracts in Monroe County, NY with Big Errors

	Single Person HH	Nonfamily HH	5+ Person HH	Households With Children
Mixed	2.4%	3.1%	36.2%	8.7%
Rent	7.7%	7.7%	46.2%	53.8%
Own	2.0%	4.1%	14.3%	0.0%

Table 6: Share of Tracts in Onondaga County, NY with Big Errors

	Single Person HH	Nonfamily HH	5+ Person HH	Households With Children
Mixed	2.6%	5.2%	37.7%	13.0%
Rent	0.0%	0.0%	50.0%	64.3%
Own	2.1%	0.0%	14.6%	2.1%

Results: Bias in the Person File

- Mean errors indicate fewer bias issues in the person file than household file
  - 4.71 more young adults were counted in rental majority areas of Onondaga County in the SF1 file than the DHC
- Rental majority areas still most prone to bias

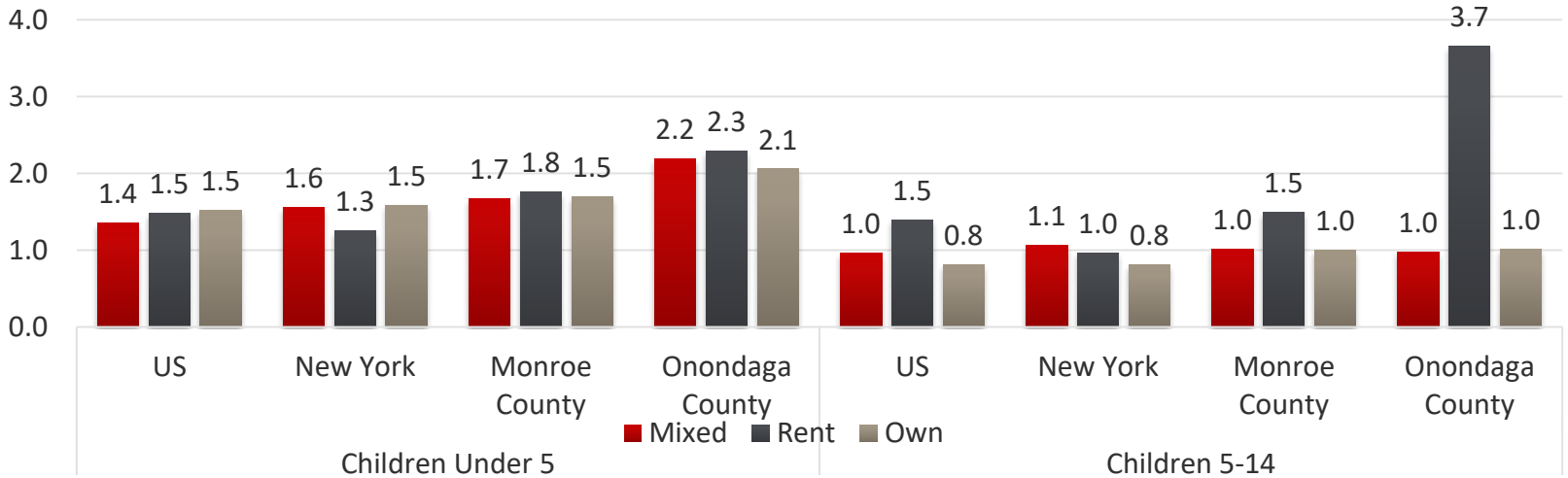
Table 7: Mean Error by Tenure Majority and Geographic Area

	Aggregate Level	Mixed	Rented	Owned
Children 0-4	United States	-0.03	**0.38***	0.00
	New York State	-0.11	0.02	0.23
	Monroe County	-0.65	2.00	0.49
	Onondaga County	-0.03	-0.21	-0.35
Children 5-14	United States	0.00	**1.18***	**-0.26***
	New York State	-0.01	0.07	-0.04
	Monroe County	-0.07	1.77	0.51
	Onondaga County	0.31	3.36	-0.31
Adults 18-24	United States	**-0.40***	**-3.27***	**0.24***
	New York State	-0.28	**-1.42***	-0.48
	Monroe County	0.11	-3.23	0.70
	Onondaga County	-0.47	-4.71	-0.02
Women 18-24	United States	**-0.16***	**-1.85***	**0.27***
	New York State	-0.06	**-0.86***	-0.07
	Monroe County	0.22	-1.69	0.38
	Onondaga County	-0.41	**-4.14***	0.63
Men 18-24	United States	**-0.24***	**-1.42***	-0.03
	New York State	**-0.22***	**-0.57***	**-0.42***
	Monroe County	-0.11	-1.54	0.32
	Onondaga County	-0.06	-0.57	-0.65

*Errors significantly different from zero are bolded and denoted by an asterisk

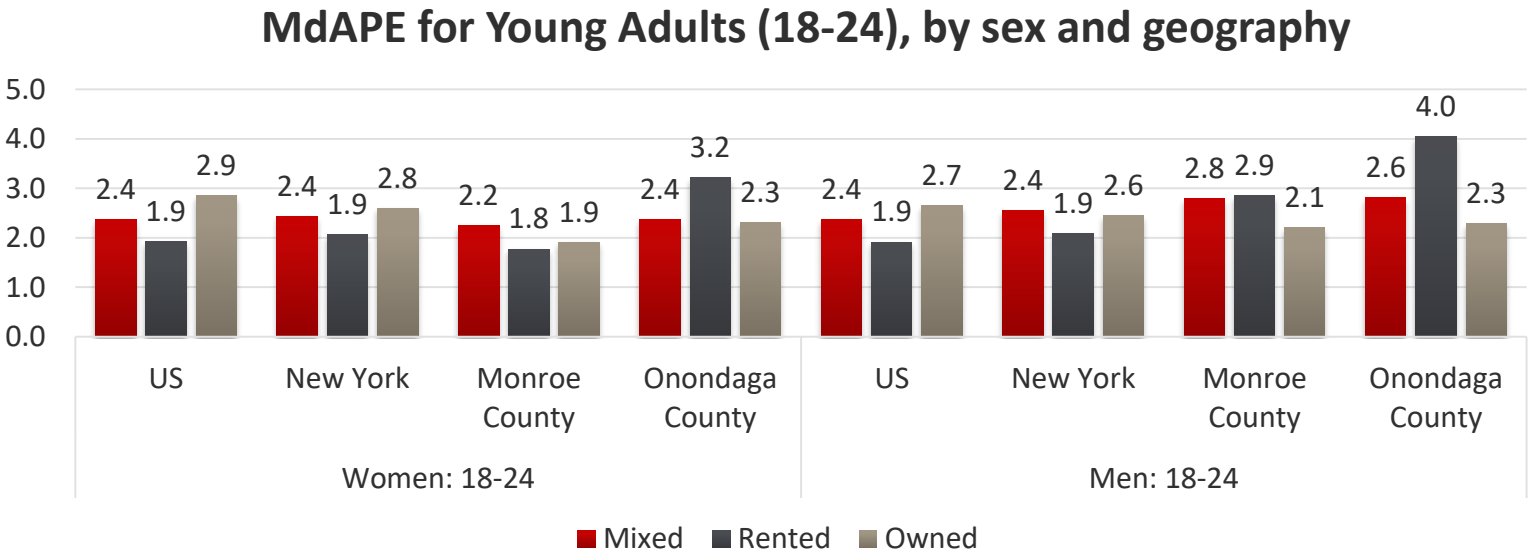
Results: Accuracy in the Person File

Figure 8: MdAPE for Children, by age group and geography



- Estimates for rental majority areas were more accurate in the person file than the housing unit file
  - The largest MdAPE for each variable was still found in rental majority areas
  - Some variation in accuracy within aggregated geographies

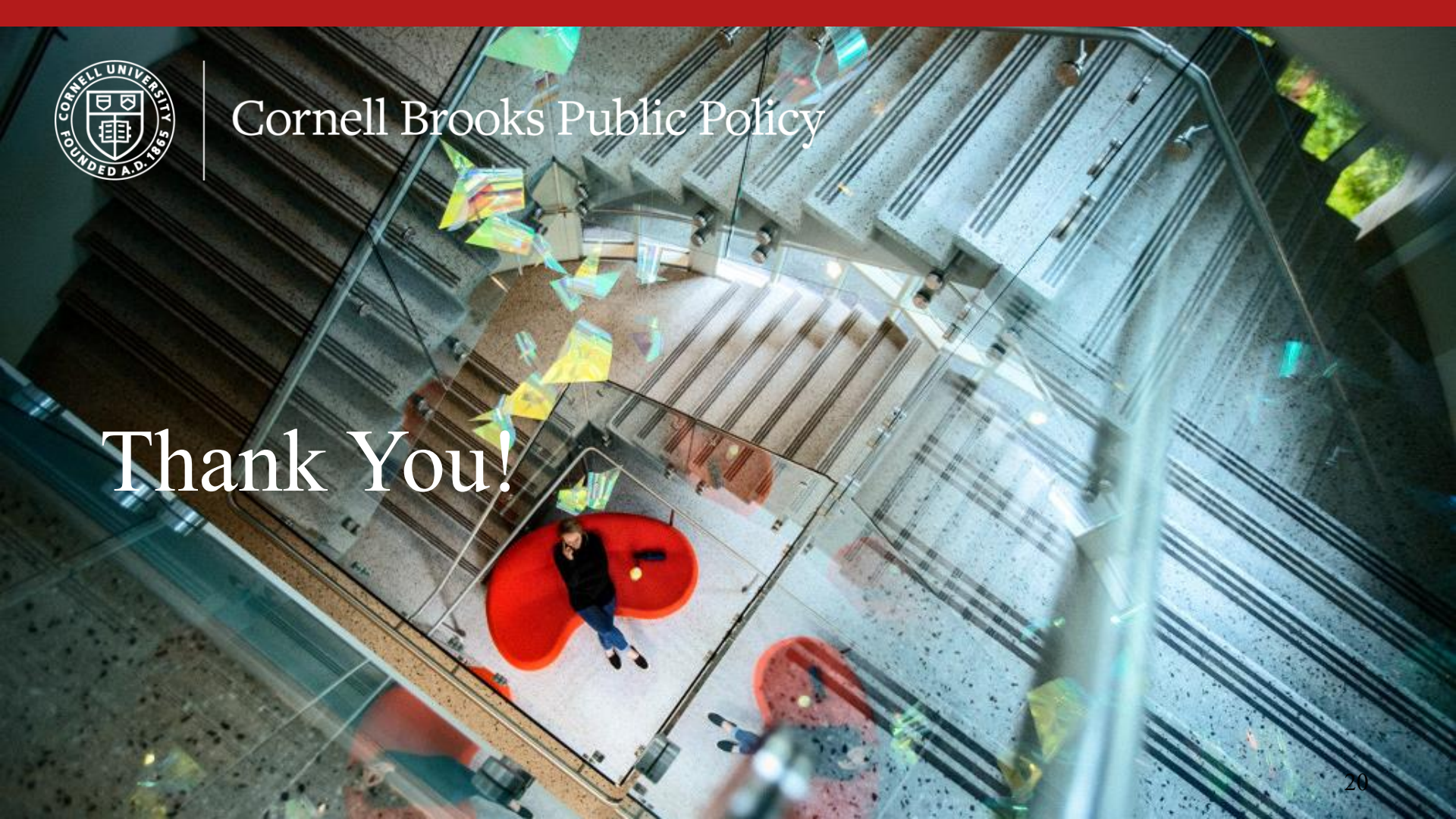
Results: Accuracy in the Person Files (cont.)



- Precision issues between the person files were limited (not shown)
  - All shares of tracts with big errors were below 5%

Conclusions

- Broad measures of similarity between SF1 and the Demonstration data can be misleading
- Tracts aggregated to sub-state levels generally produced the largest errors
- Examining metrics of error by tenure majority reveal differentials in accuracy of the DHC
    - Data for households in rental majority areas tended to be least accurate to the original 2010 summary file 1
    - Data on households in owner majority areas most accurately matched the original SF1
    - Measurement of Households with children and large households in rental majority areas contained the most bias, and had the least precision
- Issues of bias and accuracy were less prominent in the person files


Cornell Brooks Public Policy
Thank You!