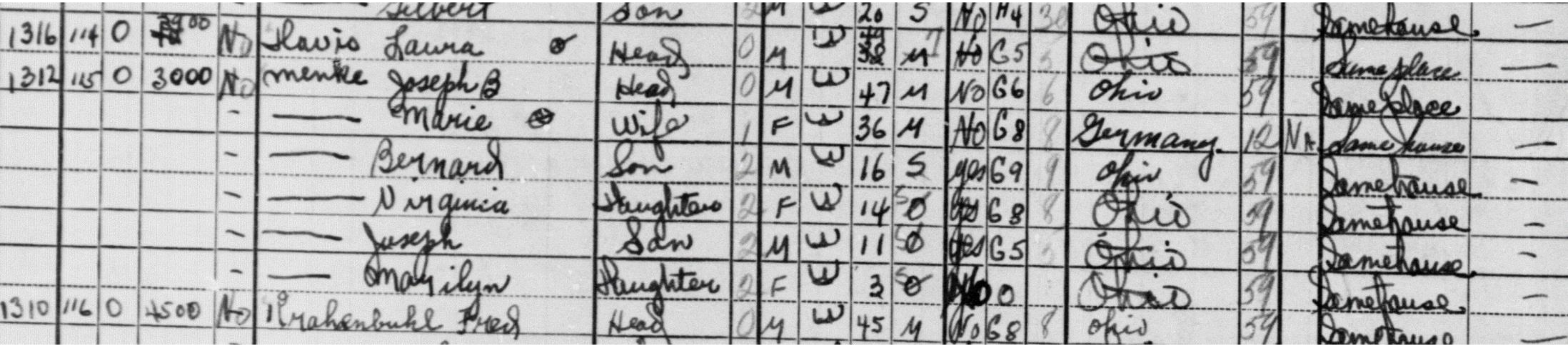


						Gilbert	Son	2	M	W	20	S	No	H4	38	Ohio	59		Same house.	—
1316	114	O	[illegible]00	No	Slavis	Laura ⊗	Head	0	M	W	32	M	No	G5	5	Ohio	59		Same place	—
1312	115	O	3000	No	Menke	Joseph B	Head	0	M	W	47	M	No	G6	6	Ohio	59		Same place	—
				–	——	Marie ⊗	Wife	1	F	W	36	M	No	G8	8	Germany	12	Na	Same house	—
				–	——	Bernard	Son	2	M	W	16	S	yes	G9	9	Ohio	59		Same house.	—
				–	——	Virginia	Daughter	2	F	W	14	S	yes	G8	8	Ohio	59		Same house	–
				–	——	Joseph	Son	2	M	W	11	S	yes	G5	5	Ohio	59		Same house	–
				–	——	Marilyn	Daughter	2	F	W	3	S	No	0		Ohio	59		Same house.	—
1310	116	O	4500	No	Krahenbuhl	Fred	Head	0	M	W	45	M	No	G8	8	Ohio	59		Same house	—

Validating Small Area Population Estimates Using Historical Census Data

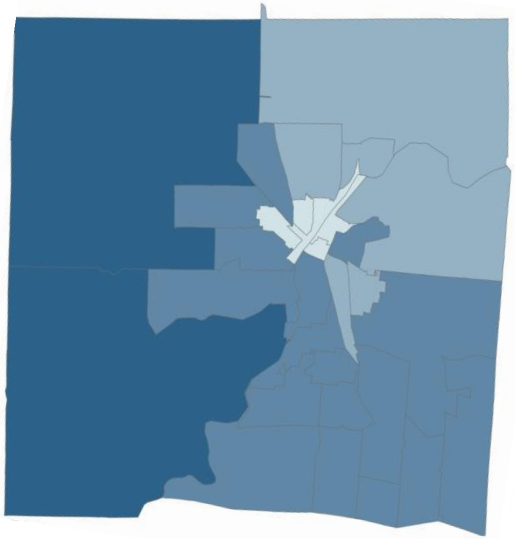
Matt Ruther, Galen MacLaurin, Stefan Leyk, Barbara Buttenfield
University of Colorado Boulder

Nicholas Nagle
University of Tennessee

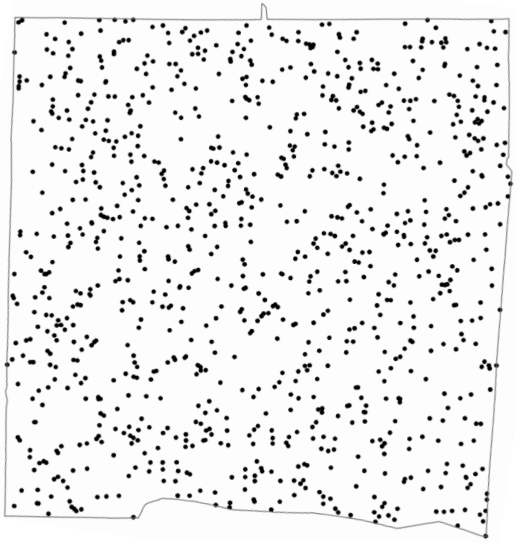
April 13, 2013

This research is funded by the National Science Foundation: “Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata”, BCS-0961598.

The Problem (What We Hope to Accomplish)



Summary Data
tracts (or other subareas)
fine geographic scale
limited demographic detail



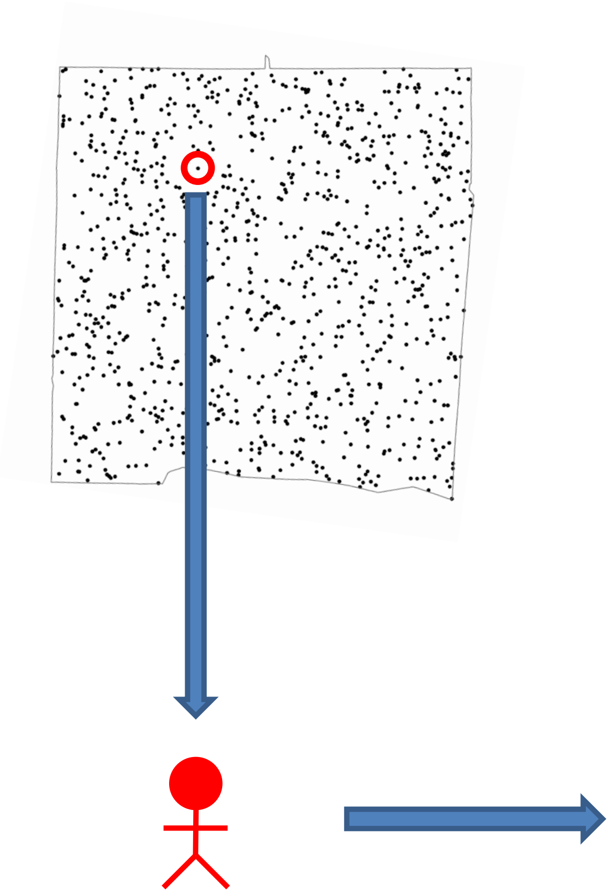
PUMS Data (microdata)
individual households
coarse geographic scale
extensive demographic detail

Spatially Allocated Microdata
fine geographic scale
extensive demographic detail

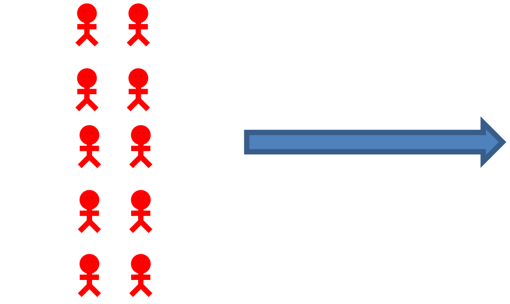
Imputation (and Allocation) in Pictures

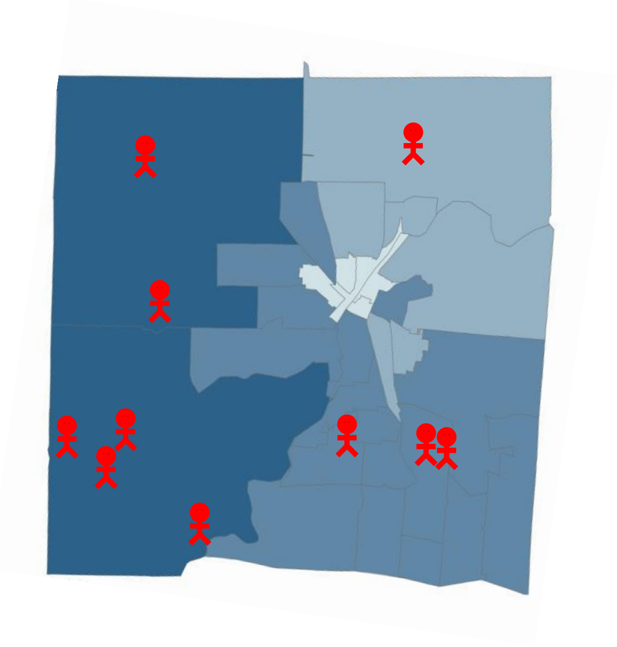
Probabilistically impute new household weights for **each** PUMS record for **each** of the tracts within the PUMA, based on the known populations of the tracts and some attributes (constraining variables) of the household.

Does not "place" individual households!



1 PUMS Record
(HH Weight=10)





Maximum Entropy Imputation

$$\text{maximize} \sum_i \sum_j w_{ij} \log\left(\frac{w_{ij}}{d_{ij}}\right)$$

$$\text{subject to} \sum_i w_{ij} x_{ik} = x_{jk} \text{ for all } j, k$$

- *i* is a household, *j* is a tract in the PUMA, *k* is an attribute
- d_{ij} is the design weight (or prior weight), w_{ij} is the imputed weight

HH #	Design Weight	Tract 1	Tract 2	Tract 3
1	7	2.33	2.33	2.33
2	16	5.33	5.33	5.33
3	14	4.66	4.66	4.66
⋮	⋮	⋮	⋮	⋮

IPF →

HH #	Design Weight	Tract 1	Tract 2	Tract 3
1	7	1.00	4.75	1.25
2	16	2.64	2.15	11.21
3	14	2.40	6.35	5.25
⋮	⋮	⋮	⋮	⋮

Benefits of the 1880 Census

- 100% count of the population publicly available (IPUMS)
- Full demographic detail and similar collection of population attributes
- Comparable spatial structure to contemporary censuses:

  State Economic Area (SEA) ≈ PUMA
  Enumeration District (ED) ≈ Census Tract

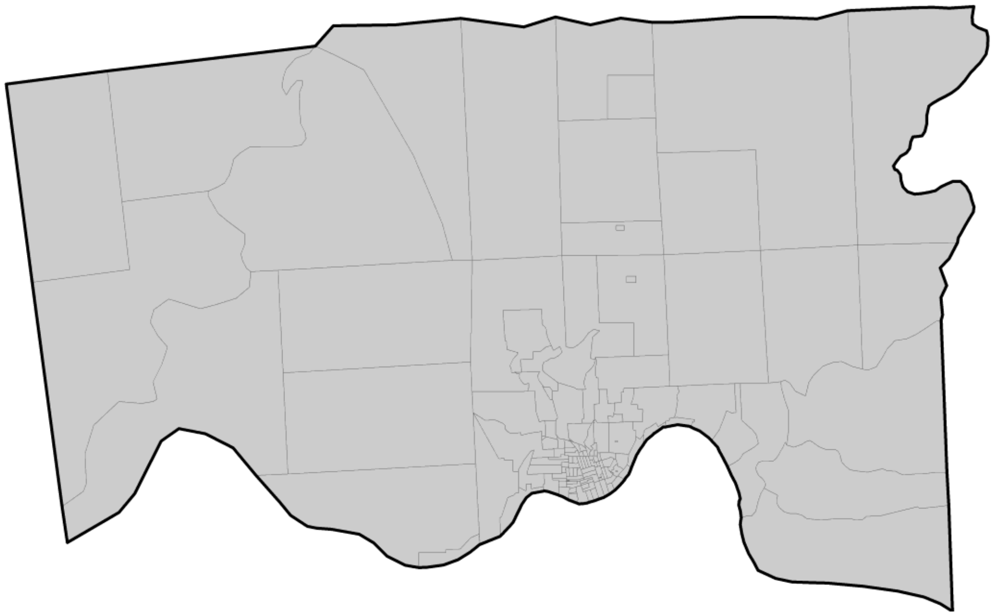
- Spatial identifiers indicating location of household

1880 Validation Goals

- How does the model perform overall?
- How can we speed up the validation when accessing confidential data at a Census Research Data Center (CDRC)?
- What types of validation can be carried out without access to confidential data at a CRDC?
- How does changing model parameters affect allocation performance?

1880 Census Geography and Data

Hamilton County, Ohio



Source: Urban Transition Historical GIS Project

1 State Economic Area
135 Enumeration Districts

Households: 68,160

Construct summary tables (for each enumeration district) from 100% microdata

Construct 5% synthetic PUMS from random sample of 100% microdata (design weight=20)

Synthetic PUMS sample: 3,408

Variables

Constraining Variables
Urban (vs. Rural)
Group quarters (vs. Non-group quarters)
White (vs. Non-white)
Foreign born (vs. Non-foreign born)
Occupation: Low-skill (vs. All other)

Validation Variables (of Household\Householder)	
Gender:	*Male*
Marital Status:	*Single, Married*
Children:	*Any Children, 5+ Children*
Age:	*0-17, 18-34, 35-49, 50+*
Nativity Status:	*Native born (2nd Gen)*
Farm Status:	*Farm*

Evaluating Error

	Allocated	Actual	Residual
Enumeration District 1	110	110	0
Enumeration District 2	152	150	2
Enumeration District 3	127	140	13
Total	389	400	11

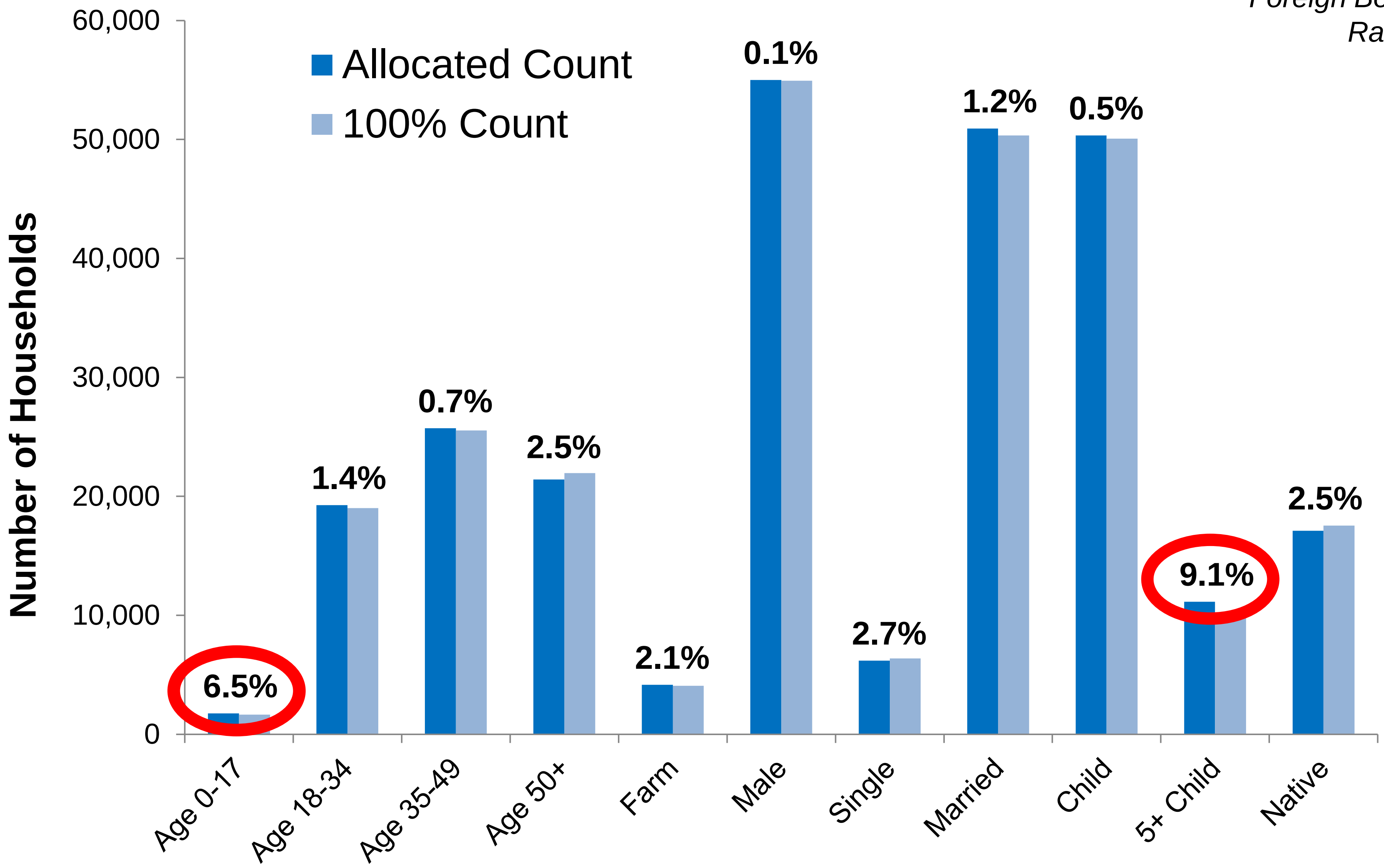
$$\boldsymbol{Error\ in\ Margin} = \left|\frac{Residual\ Total}{Actual\ Total}\right| = \left|\frac{389-400}{400}\right| = 0.03$$

$$\boldsymbol{Allocation\ Error\ (ED_i)} = \left|\frac{Residual\ ED_i}{Actual\ ED_i}\right| = \left|\frac{13}{140}\right| = 0.09$$

$$\boldsymbol{Total\ Allocation\ Error = TAE} = \frac{\sum_i |Residual\ ED_i|}{Actual\ Total} = \frac{|15|}{400} = 0.04$$

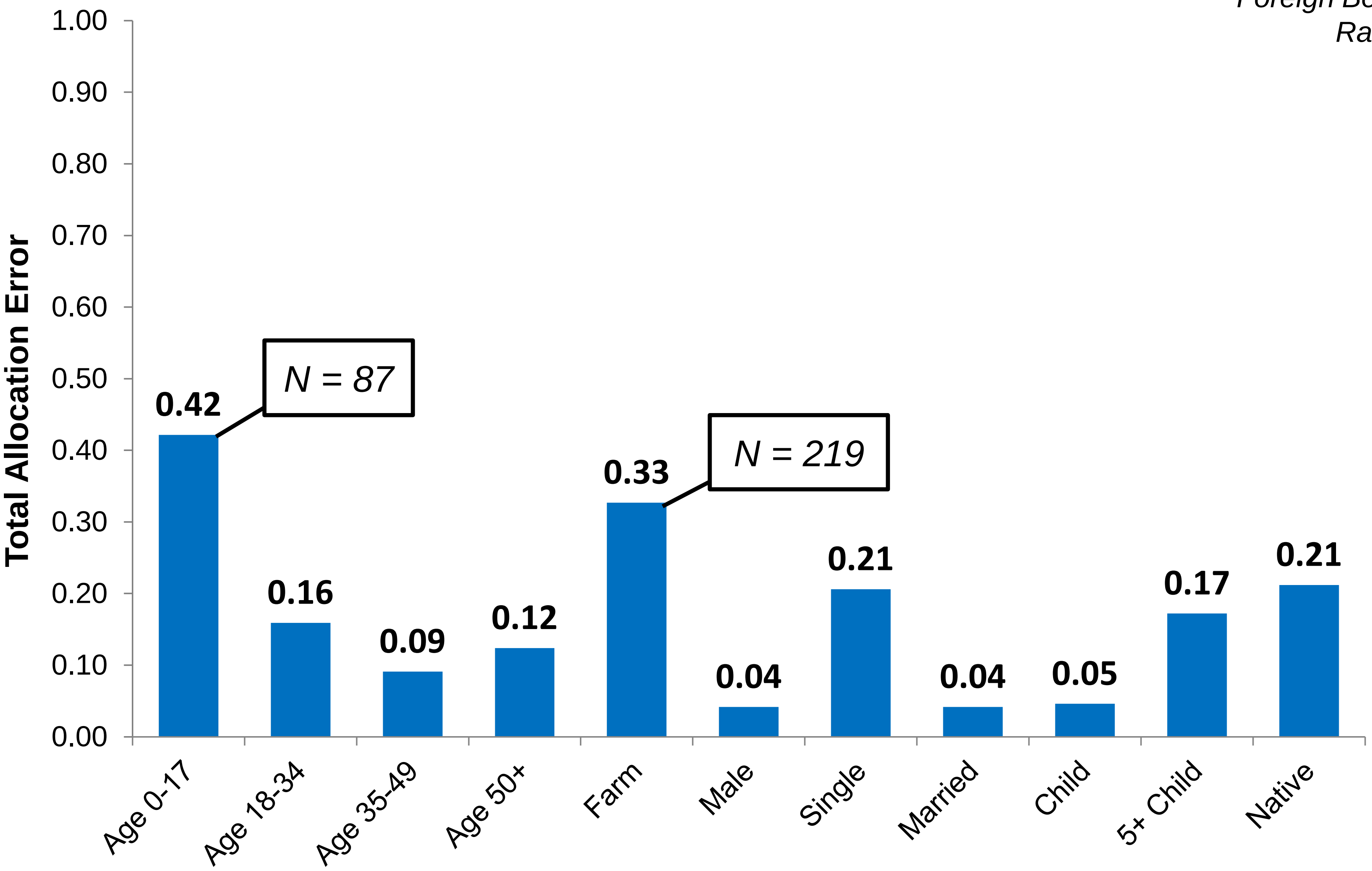
Error in Margin

Constraints
Urban/Rural
Group Quarters
Occupation
Foreign Born
Race



Total Allocation Error (TAE)

Constraints
Urban/Rural
Group Quarters
Occupation
Foreign Born
Race

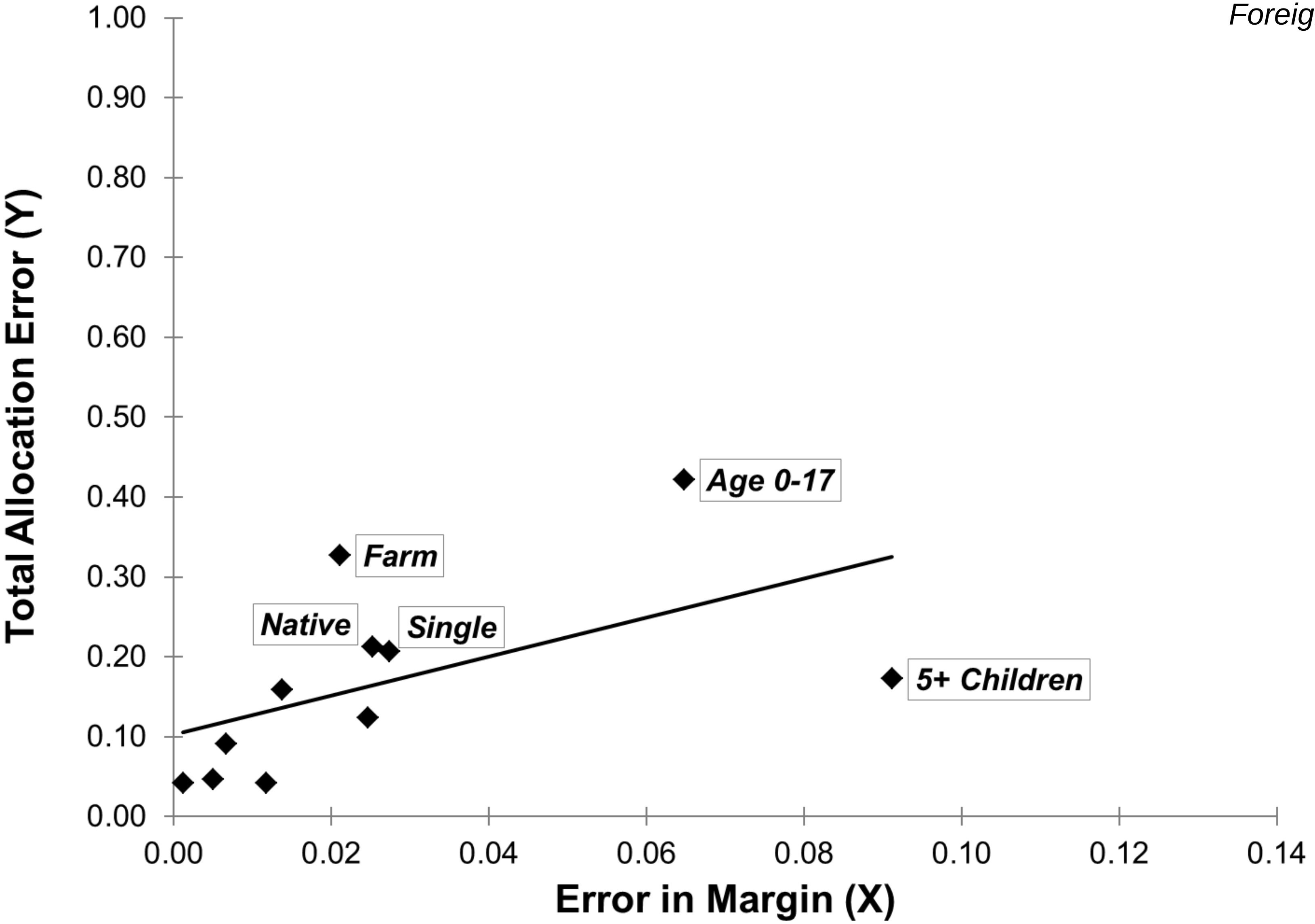


N = 87

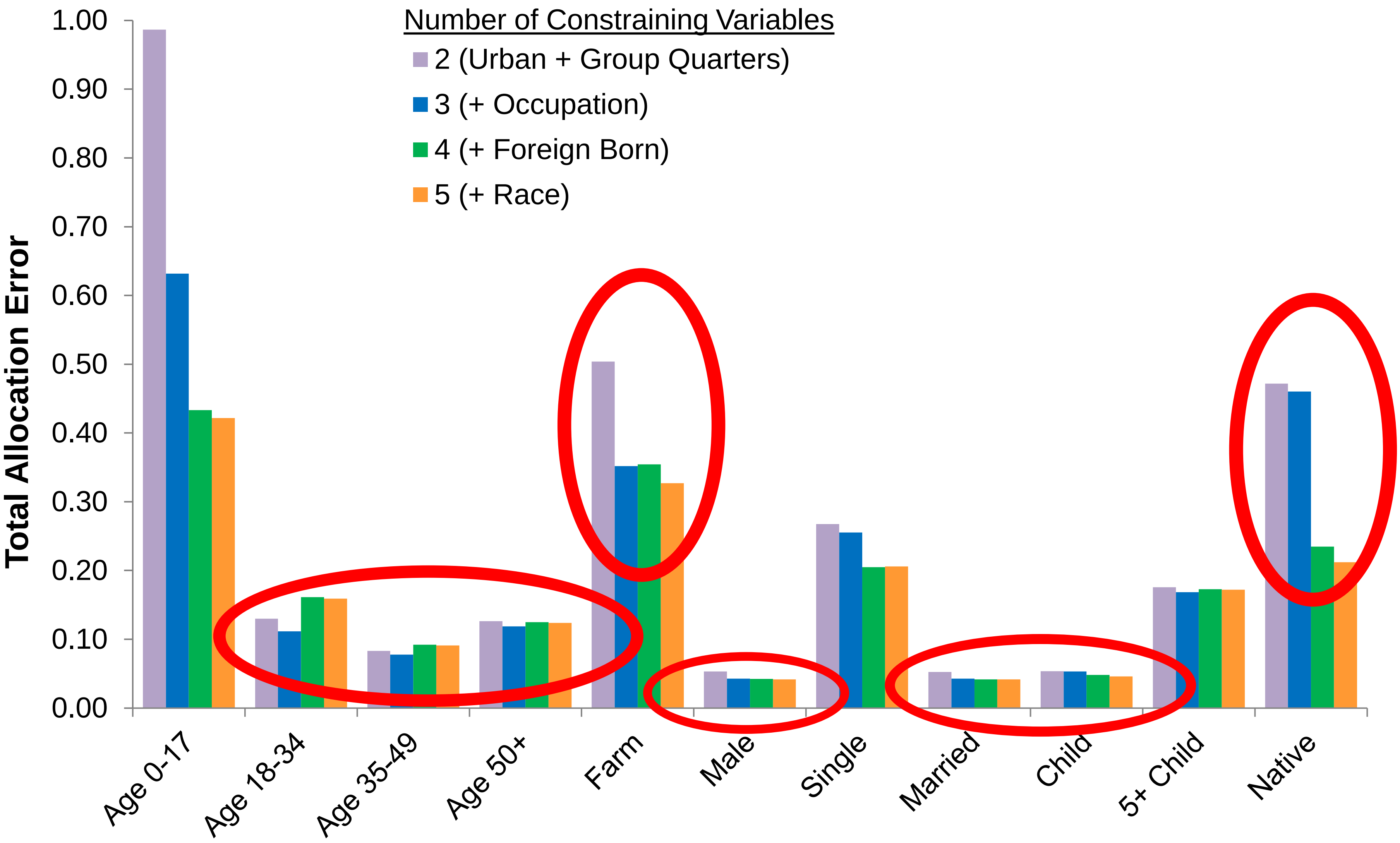
N = 219

Error in Margin and TAE

Constraints
Urban/Rural
Group Quarters
Occupation
Foreign Born
Race

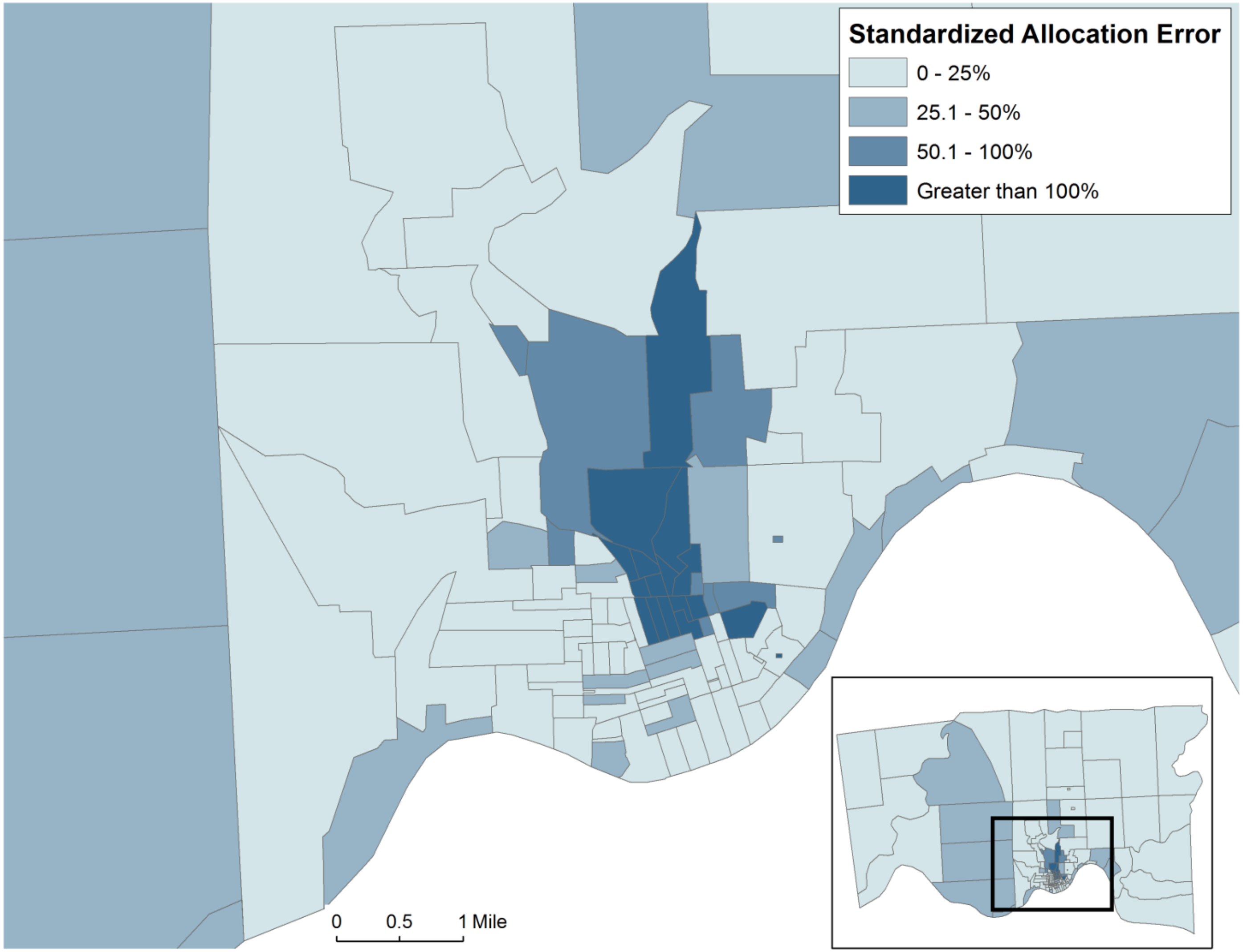


Total Allocation Error: Comparing Models



Spatial Heterogeneity in Allocation Errors:

2nd Generation Native Born Households



Validation Conclusions

How does the model perform overall?

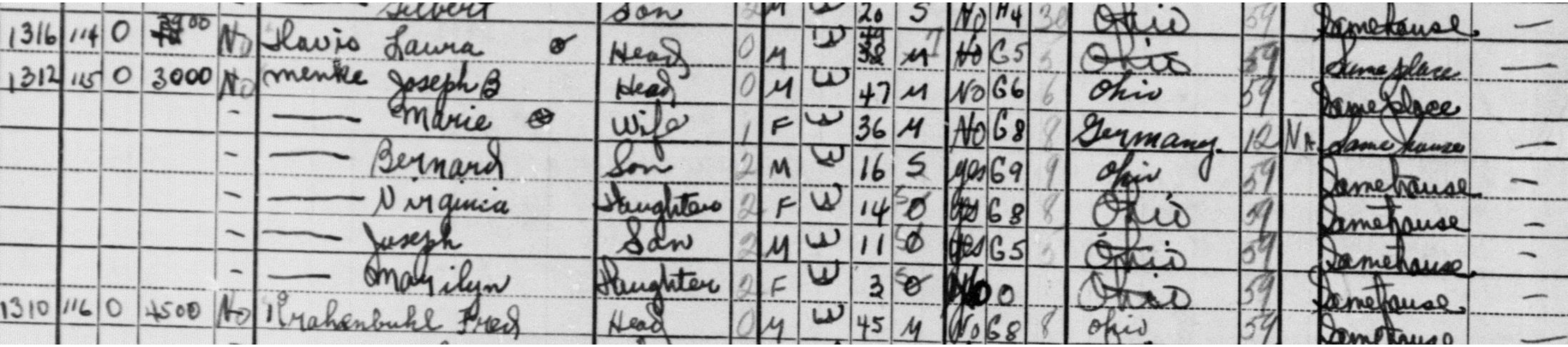
- Initial allocation results are promising

How can we streamline the validation prior to accessing confidential data at a CRDC?

- Much of this procedure can be carried out prior to visiting CRDC
- Compare metrics for variables available in summary tables

How does changing parameters affect performance?

- Generally, additional constraints improve TAE
- Additional constraint show notable improvement on variables with which they are correlated



						Gilbert	Son	2	M	W	20	S	No	H4	38	Ohio	59		Samehouse.	—
1316	114	O	3900	No	Slavis	Laura ⊗	Head	0	M	W	39	M	~~No~~	G5	5	~~Ohio~~	59		Same place	—
1312	115	O	3000	No	Menke	Joseph B	Head	0	M	W	47	M	No	G6	6	Ohio	59		Same place	
				–	———	Marie ⊗	Wife	1	F	W	36	M	~~No~~	G8	8	Germany.	12	Na	Same house	—
				–	———	Bernard	Son	2	M	W	16	S	yes	G9	9	Ohio	59		Samehouse.	—
				–	———	Virginia	Daughter	2	F	W	14	0	yes	G8	8	Ohio	59		Samehouse	—
				–	———	Joseph	Son	2	M	W	11	0	yes	G5	5	Ohio	59		Samehouse.	—
				–	———	Marilyn	Daughter	2	F	W	3	0	No	0		Ohio	59		Samehouse.	—
1310	116	O	4500	No	Krahenbuhl	Fred	Head	0	M	W	45	M	No	G8	8	Ohio	59		Samehouse	—

Matt Ruther
matthew.ruther@colorado.edu
Department of Geography
University of Colorado Boulder

This research is funded by the National Science Foundation: “Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata”, project BCS-0961598 awarded to University of Colorado at Boulder and University of Tennessee at Knoxville. Funding by NSF is gratefully acknowledged.