Exploratory Imputation Research Results Using Data from Schools and Staffing Survey and Common Core of Data

Sarah Konya (presenter)
Jacob Enriquez, Mei Li, Svetlana Mosina,
T. Trang Nguyen, Allison Zotti
(coauthors)
U.S. Census Bureau

Outline

- Background of Survey
- Research Objective 1: Comparison of Administrative Data to Survey Data
- Research Objective 2: Comparison of Imputation Methods
- Research Objective 3: Matching Variables Analysis

Background of Survey

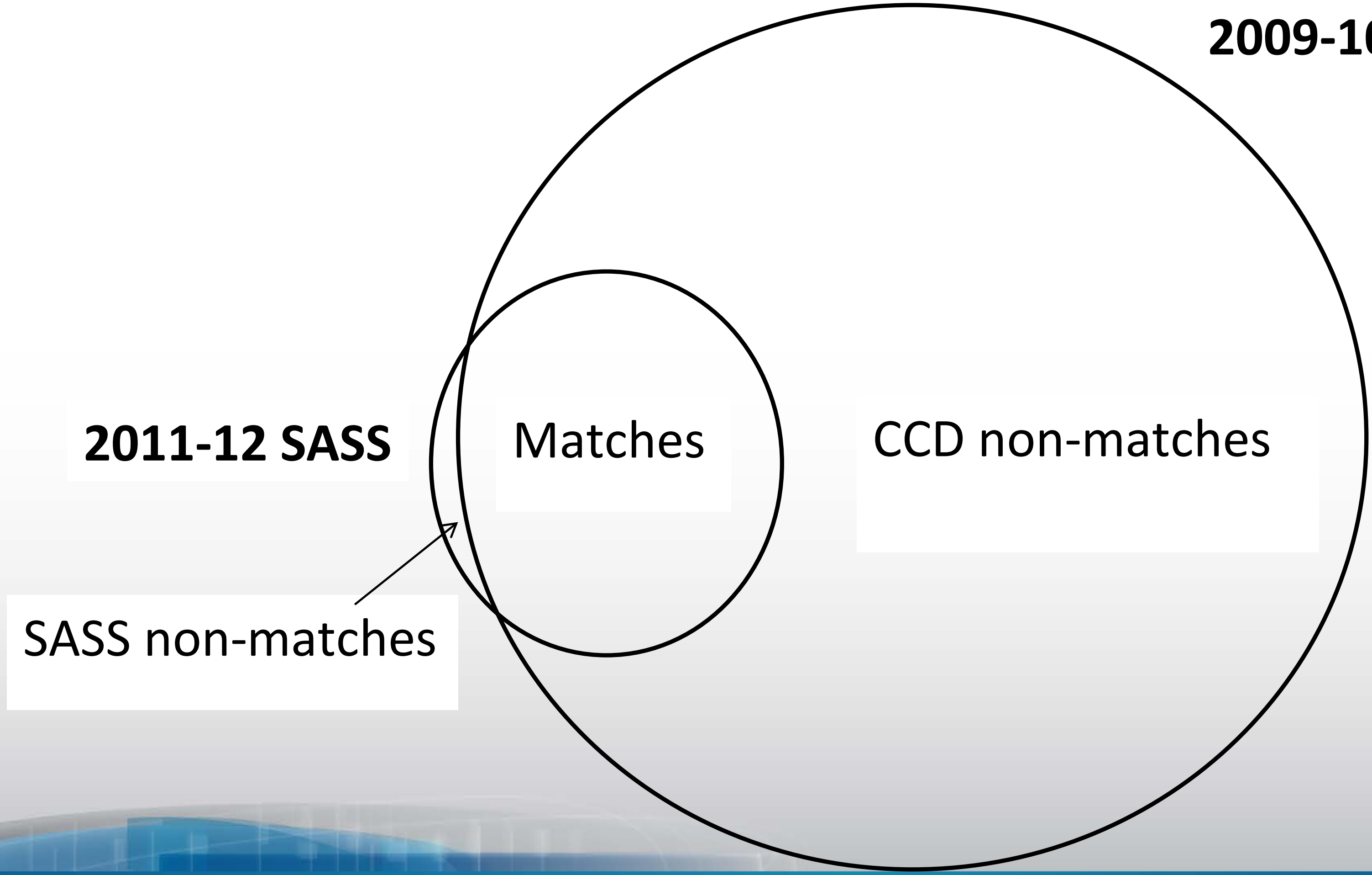
- Schools and Staffing Survey (SASS)
  - American elementary and secondary education
  - Sponsor: National Center for Education Statistics
  - Conducted every four years
- Frame built from the Common Core of Data (CCD) administrative data file
  - Approx. 100,000 schools on CCD
  - Approx. 10,000 schools sampled for SASS
- Redesign: National Teacher and Principal Survey (NTPS)
  - Conducted every two years

Research Objective 1

- Could the data from the CCD potentially be used to completely replace SASS items where response information is also available from the CCD?

Item Description	Type of Question
Grades offered (15)	Binary
Total enrollment	Discrete
Enrollment by race (8)	Discrete
School type	Ordinal
Teacher count (3)	Discrete

Administrative Records Coverage Results: School-level Matching Rate



2009-10 CCD

2011-12 SASS

Matches

CCD non-matches

SASS non-matches

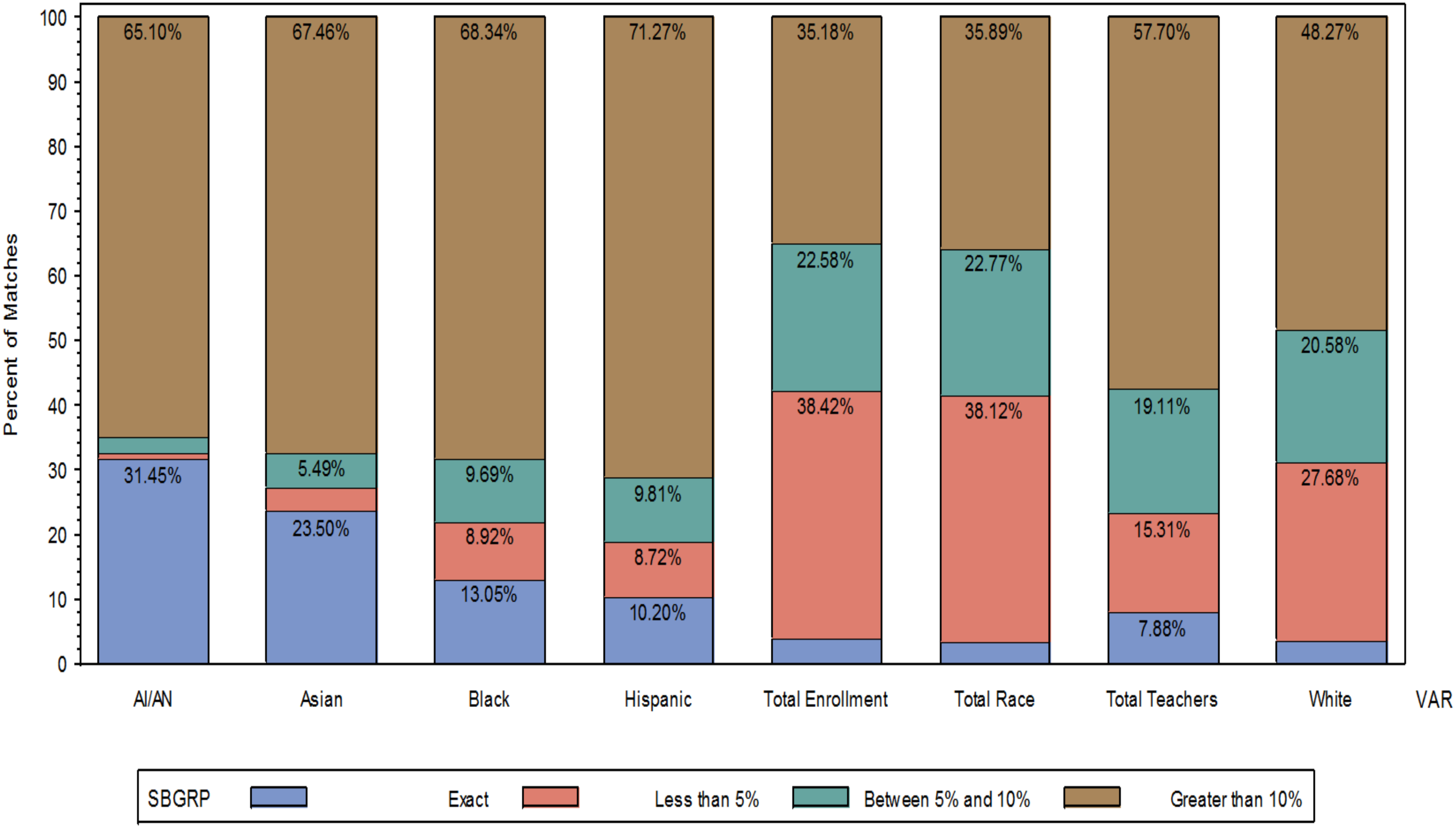
- Matched the two datasets on unique school ID
- 10,614 of the 11,000 SASS records matched
  - 96.5% match

Administrative Records Coverage Results: Rate of Reported CCD Values

$$= \frac{Number\ of\ Schools\ on\ the\ CCD\ with\ a\ Nonmissing\ Value}{Total\ Number\ of\ Matching\ Schools} * 100$$

Item Description	Rate of Reported values on CCD
School type	100%
Total enrollment	98.41%
Hispanic enrollment	98.31%
White enrollment	98.31%
Black enrollment	98.31%
Asian enrollment	98.31%
American Indian/ Alaskan Native enrollment	98.31%
Total ethnicity enrollment	98.31%
Total teachers	97.84%

Relative Differences between 2009-10 CCD and 2011-12 SASS Values



Paired T-Test between 2011-12 SASS and 2009-10 CCD Values

Item Description	Mean difference	Std. Dev.	N	t Value	Pr > |t|
Total enrollment	-24.85	221.7	7,109	-9.45	< .0001
Hispanic enrollment	-9.48	93.56	6,676	-8.28	< .0001
White enrollment	-8.37	147.9	6,662	-4.62	< .0001
Black enrollment	-0.37	52.22	6,674	-0.58	0.5671
Asian enrollment	1.58	29.83	6,682	4.33	< .0001
American Indian/ Alaskan Native enrollment	-0.73	14.94	6,696	-4.02	< .0001
Total ethnicity enrollment	-29.89	221.50	7,102	-11.37	< .0001
Total teachers	-4.45	13.46	7,068	-27.79	< .0001

Objective 1 Conclusion

- Overall coverage of the CCD was good
  - 96.5% of SASS records matched to CCD
- Coverage of reported CCD values was good for 9 items
- Black enrollment is the only item where SASS and CCD values were not significantly different
- More research should be done using multiple years of CCD and SASS data to make a decision on the replacement of SASS data

Research Objective 2

- Should the hot deck imputation method for SASS be replaced with a multiple imputation method?

Item Description	Response Rate Percent	Type of Question
Black Enrollment	93.84	Discrete
Pension Check (how much)	72.51	Continuous

Current Imputation Method

- Consistency edits
- Logic edits
- SASS hot deck imputation

Advantages and Disadvantages of Hot Deck Imputation

Advantages	Disadvantages
Intuitively simple method	Donor selected may not be similar to the record to be imputed
No distributional assumptions on the data	Using the same donor too many times
Does not rely on model fitting	May yield biased estimates and underestimate standard errors

Alternative Multiple Imputation Methods

Imputation Method	Description
Markov Chain Monte Carlo (MCMC)	Arbitrary missing pattern, generates pseudorandom draws from probability distributions via Markov chains, imputes with model-produced values
Propensity Score	Monotone missing pattern, conditional probability to assign value to imputed item using regression, imputes with donor values
Regression	Monotone missing pattern, fitting a model that relates the response variable to the covariates, imputes with model-produced values
Predictive Mean Matching (PMM)	Monotone missing pattern, linear prediction as a distance measure for the set of nearest neighbors (donors) consisting of the complete values, the respondent with the smallest distance metric is chosen as the donor, imputes with donor values

Imputation Model Covariates

Items to Impute	Covariates	Adj. R^2
Black Enrollment	CCD Black Enrollment, Total Teachers, CCD Free and Reduced Lunch, Number of Vice Principals, Number of Black Teachers	0.8905
Pension Check	Highest Degree Attained by Teacher, Number of Years as a Teacher	0.0934

Evaluation Measures for Black Enrollment

Method	Avg. of MRD	Avg. of Q1 Bias	Avg. Median Bias	Avg. of Q3 Bias	Avg. of Relative Bias	Avg. of Mean Bias	Avg. Std. Dev. Bias	% of Datasets T-Test was sig.
MCMC	5.67	36.24	31.40	4.26	0.23	19.08	-17.46	100.00
Propensity	8.67	33.30	41.02	11.50	0.00	-0.09	-87.27	0.84
PMM	0.33	0.63	0.23	0.88	0.00	-0.06	-2.35	6.30
Regression	5.67	36.26	31.38	4.52	0.23	19.09	-17.40	100.00

Evaluation Measures for Pension Check

Method	Avg. of MRD	Avg. of Q1 Bias	Avg. Median Bias	Avg. of Q3 Bias	Avg. of Relative Bias	Avg. of Mean Bias	Avg. Std. Dev. Bias	% of Datasets T-Test was sig.
MCMC	10.46	18478	6862.8	-4479.1	0.24	4502.3	-13592	94.40
Propensity	7.94	12053	2236.3	-7400.4	0.00	-21.67	-11786	10.00
PMM	5.61	559.2	-1835.5	1298.2	0.01	58.45	-214.95	20.80
Regression	10.49	18503	6901.8	-4430.9	0.24	4533.6	-13587	94.00

Comparing PMM to Hot Deck

- Chose PMM as best alternative method

- Compare to:
  - SASS Hot Deck
  - Common Hot Deck

T-Test of the Means

Data	Statistic	Black Enrollment	Pension Check
Reported	Mean	85.10	19398.10
	Std. Dev.	159.30	18166.90
	N	7020	670
Imputed with PMM	Mean	123.30	14820.60
	Std. Dev.	208.60	16874.80
	N	416	235
	p-value	0.0003	0.0005
Imputed with SASS Hot Deck	Mean	99.09	20534.50
	Std. Dev.	168.20	22959.70
	N	461	233
	p-value	0.0831	0.4940
Imputed with Common Hot Deck	Mean	100.3	18000.20
	Std. Dev.	179.70	15734.70
	N	461	235
	p-value	0.0763	0.5891

Objective 2 Conclusion

- PMM is the best alternative imputation method for the items researched
- SASS hot deck and Common Hot Deck methods better at preserving the means of data than PMM

Research Objective 3

- Should SASS matching variables be updated if we continue to use the current hot deck method?

SASS Item	Current Matching Variables	Model Covariates
Black Enrollment	Urban Status, Minority Enrollment Code, State Group, and State	CCD Black Enrollment, CCD Free and Reduced Lunch, Total Full-time or Part-time Teachers, Total Full-time Vice/Assistant Principals, Total Black Full-time or Part-time Teachers
Newly Hired Teachers	Urban Status, Instructional Level of School, School Type, State Group, and State	White Enrollment, Black Enrollment, Total Teachers, Total Vice Principals, Number of Custodial and Security, Total Students with IEP because of Special Needs

Multiple Correlations

- Show how well the response variable can be predicted using a linear function of independent variables

		Association with Outcome	
		Low	High
Association with Non-response	Low	I Bias: Unchanged Variance: Unchanged	II Bias: Unchanged Variance: Decreases
	High	III Bias: Unchanged Variance: Increases	IV Bias: Decreases Variance: Decreases

Association with Outcome

SASS Item	Multiple Correlation using Matching Variables	Multiple Correlation using Model Covariates
Black Enrollment	0.6139	0.9437
Newly Hired Teachers	0.3603	0.5118

Association with Nonresponse

SASS Item	Number Missing	Number Reported	Multiple Correlation using Matching Variables	Multiple Correlation using Model Covariates
Black Enrollment	461	6,517	0.2782	0.0374
Newly Hired Teachers	236	6,147	0.1233	0.0490

Objective 3 Conclusion

- Neither option produced a high association with nonresponse of the SASS item
- The model covariates, overall, had a high association with the outcome of the SASS items
- Not feasible to create a unique set of covariates for each imputation item if the variance could only *potentially* be decreased

Contact information:

sarah.dial@census.gov

U.S. Census Bureau
Washington, DC 20233