Westat

Variables Appended to ABS Frames: Has Data Quality Improved?

Shelley Brock Roth, Andrew Caporaso, Jill DeMatteis
Westat

AAPOR 2018 Taking Survey and Public Opinion Research to New Heights

Overview

- Introduction
  - Research goals
  - Description of studies used for analysis and ABS frame
  - Other relevant research
- Results
- Conclusions and Recommendations

Introduction

Research Goals

1. Evaluate availability and quality of variables appended to an ABS frame
2. Determine whether appended variables are potentially useful for oversampling
3. Investigate the potential for using appended variables for weighting adjustments for nonresponse

NHTS, HINTS, and NHES

- National Household Travel Survey (NHTS 2017)
  - Two-phase national household ABS with oversampling in some geographic areas
  - 929,077 households sampled
  - 130,000 completed surveys (Screener AAPOR RR3=30%, Extended AAPOR RR2=52%)
- Health Information National Trends Survey Cycle 1 (HINTS5 2017)
  - Single-phase national household ABS with oversampling of high minority areas
  - 13,360 households sampled
  - 3,335 completed surveys (AAPOR RR2 = 32%)
- National Household Education Survey Field Test (NHES 2011)
  - Two-phase national household ABS
  - 41,260 households sampled
  - 5,590 completed surveys (Screener AAPOR RR4=69%, Extended AAPOR RR2=73%)

ABS Frame

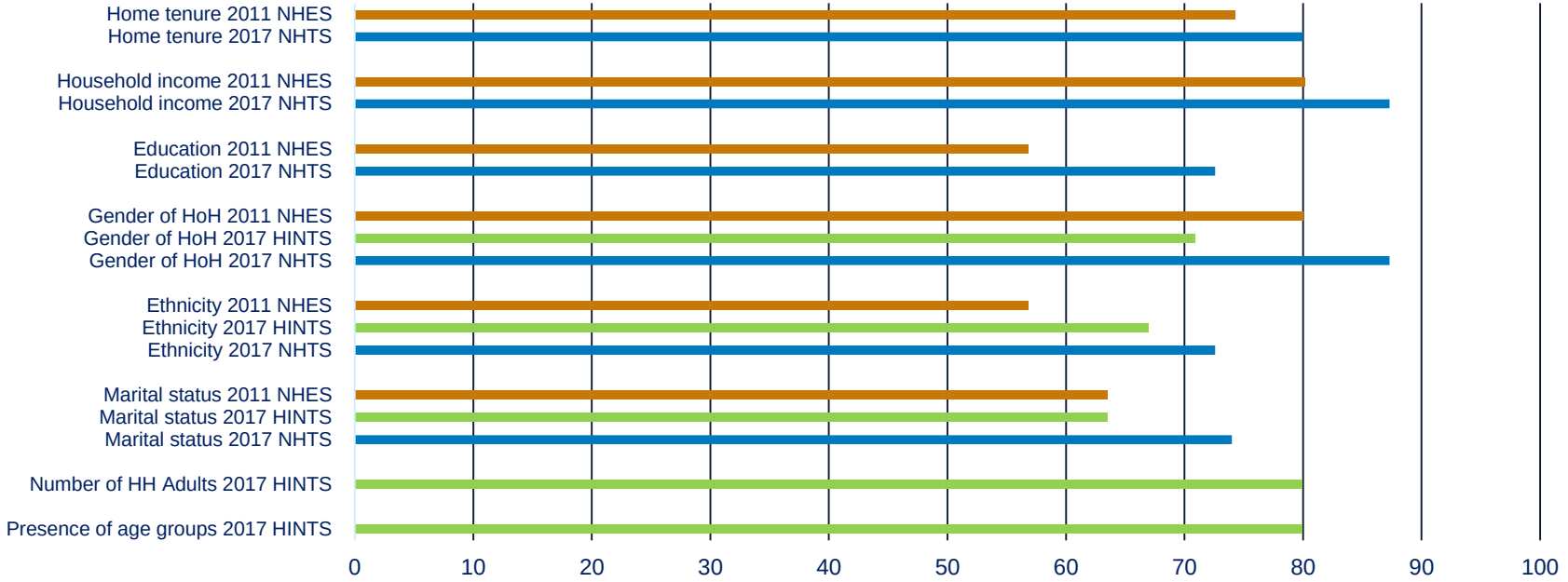
- ABS frame constructed from US Postal Service Computerized Delivery Sequence File
  - Contains basic set of postal service variables
  - Variety of additional demographic and socio-economic variables can be appended
- MSG (Marketing Systems Group)
  - Vendor who maintains ABS frame from which both samples were drawn
  - Frame updated monthly

Other Relevant Research

- Yan et al, 2011: predicting eligible household units using appended data
- Roth et al, 2013: using appended data for stratification or oversampling
- Buskirk et al, 2014: append rates of vendor data and consistency of appended data from different vendors
- English et al, 2014: enhancement of survey efficiency using targeted lists
- McMichael et al, 2014: optimal allocation for Hispanic populations based on Hispanic flags
- Valliant et al, 2014: sample stratification
- West et al, 2015: comparing two commercial data sources to NSFG and each other for survey operations and estimation

Results

GOAL1: Examine Availability
Non-Missing Rates for Appended Demographics



Key: NHES; NHTS; HINTS

GOAL 1: Examine Quality
Agreement Statistics

- Example: calculation of agreement statistics for NHTS Ethnicity

MSG Ethnicity	NHTS Ethnicity		Total	True predictivity	Overall concordance
	Hispanic	Not Hispanic			
Identified as Hispanic	6% (a)	2% (b)	9% (a+b)	72% (a)/(a+b)	91% (a+d)
Not identified as Hispanic	7% (c)	85% (d)	91% (c+d)	93% (d)/(c+d)	

GOAL 1: Examine Quality
Agreement Statistics (cont.)

NHTS

Characteristic	True +	True -	Overall Concordance
Hispanic ethnicity	**0.72**	**0.93**	**0.91**
Hispanic surname	**0.79**	**0.90**	**0.90**
Presence of children	0.76	0.58	0.73
Home is rented	**0.83**	**0.77**	**0.79**
Education HS or less	0.37	0.84	0.71
Income <$35K	0.66	0.73	0.70

HINTS

Characteristic	True +	True -	Overall Concordance
Hispanic ethnicity	**0.63**	**0.95**	**0.92**
Hispanic surname	**0.68**	**0.95**	**0.93**
18-24 present	0.44	0.92	0.88
35-64 present	0.80	0.60	0.70
25-34 present	0.32	0.88	0.80
65+ present	**0.78**	**0.81**	**0.80**
Married HH	0.72	0.66	0.69
1 adult HH	0.46	0.77	0.67
2 adult HH	0.64	0.52	0.55
2+ adult HH	0.81	0.46	0.65
3+ adult HH	0.27	0.87	0.72
Female HoH	0.75	0.48	0.55

GOAL 2: Examine Potential for Oversampling
Variables Investigated

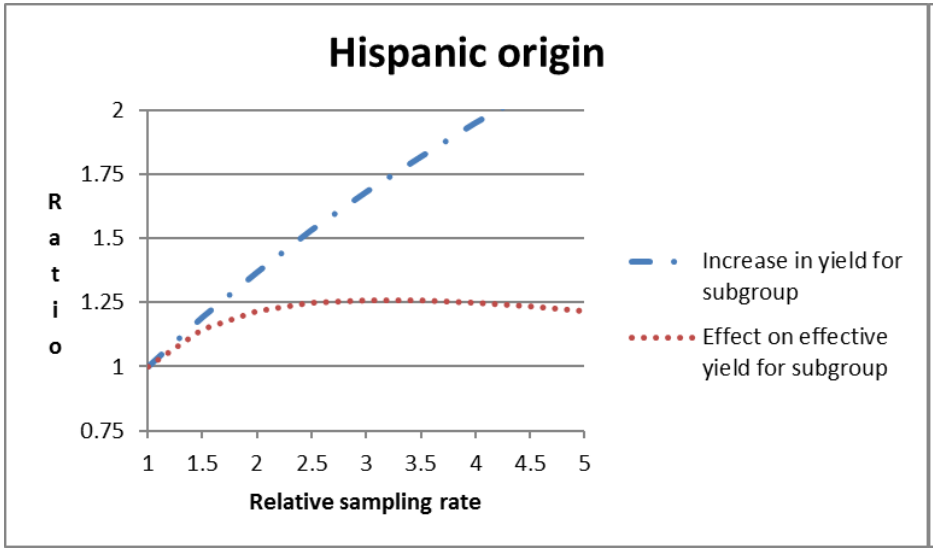
Characteristic	NHTS	HINTS
Ethnicity	X	X
Hispanic surname	X	X
Home tenure (rent, other)	X	
Educational attainment (HS or less, other)	X	
Household income (<$35K annually, other)	X	
Presence of children	X	
Number of adults = 1		X
Number of adults = 2 or more		X
Number of adults = 3 or more		X
Presence of 18-24 year old		X
Presence of 25-34 year old		X
Presence of 35-64 year old		X
Presence of 65+ year old		X
Marital status (married, not married)		X
Female Head of Household		X

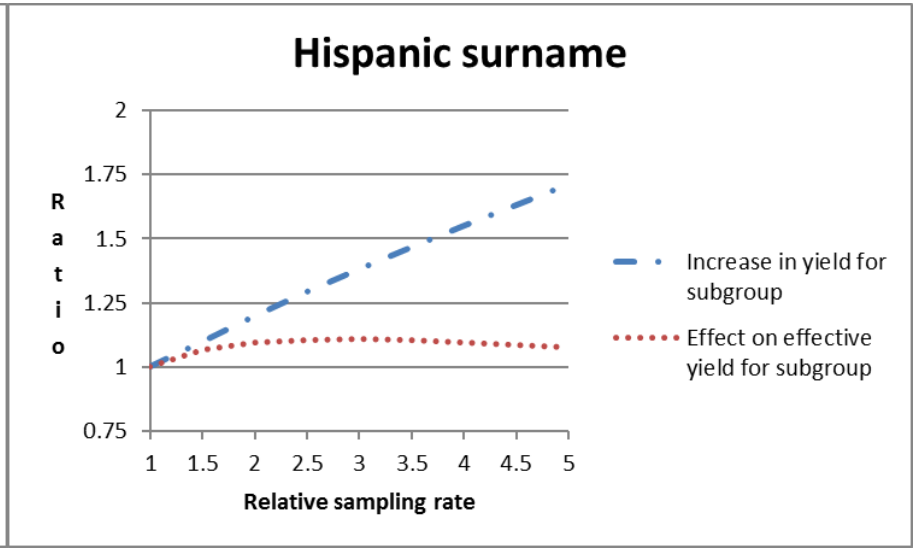
GOAL 2: Examine Potential for Oversampling Methods

- Two measures computed for each characteristic
  1. Increase in *nominal* yield for the subgroup of interest
  2. Effect of the oversampling on the *effective* yield for the subgroup of interest (accounts for design effect due to oversampling and misclassification)
- Oversampling scenarios considered two strata for each characteristic
  1. Presence of characteristic
  2. Absence of characteristic (includes missing)

Goal 2: Examine Potential for Oversampling
Good candidates

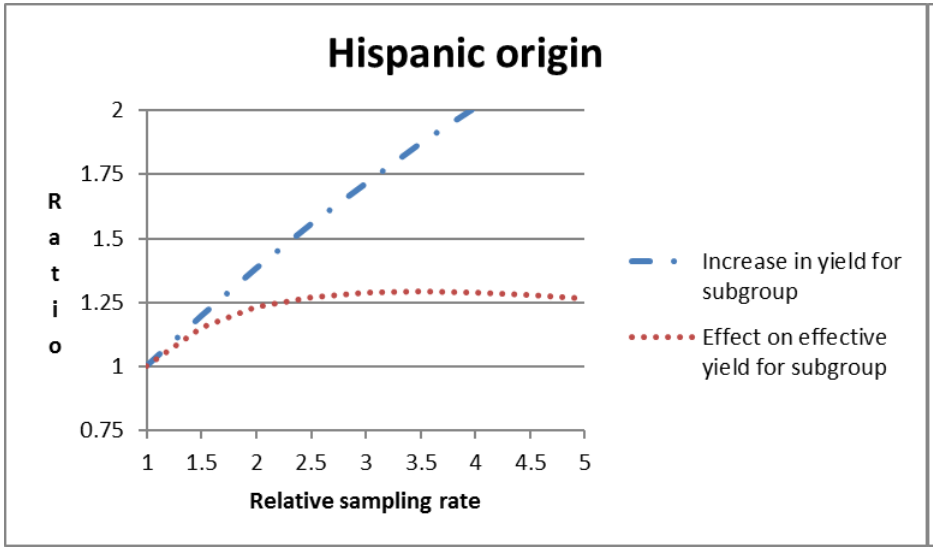
NHTS: Ethnicity and Hispanic surname

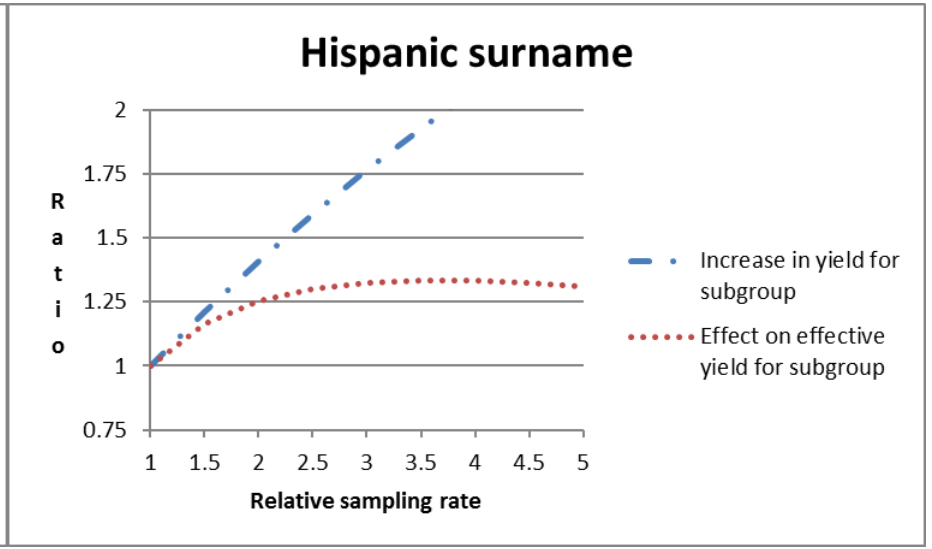




Goal 2: Examine Potential for Oversampling
Good candidates

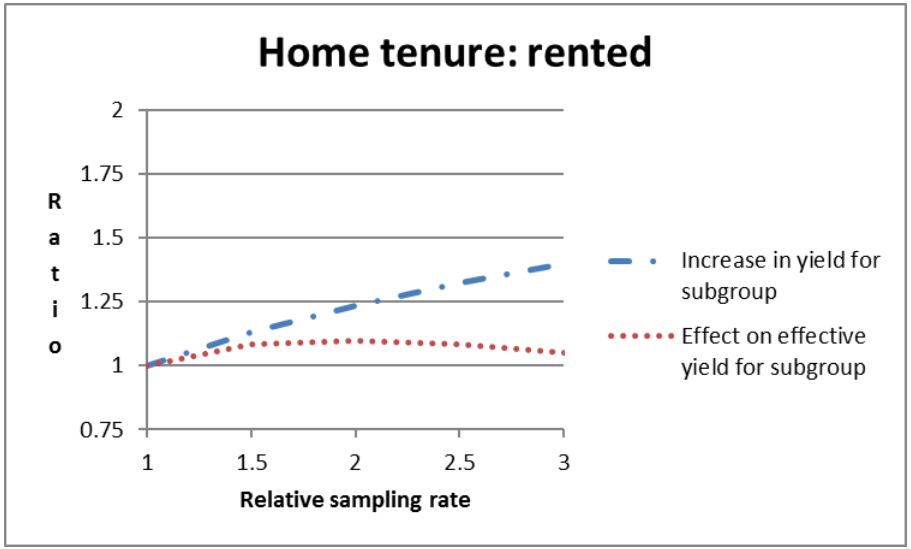
HINTS: Ethnicity and Hispanic surname





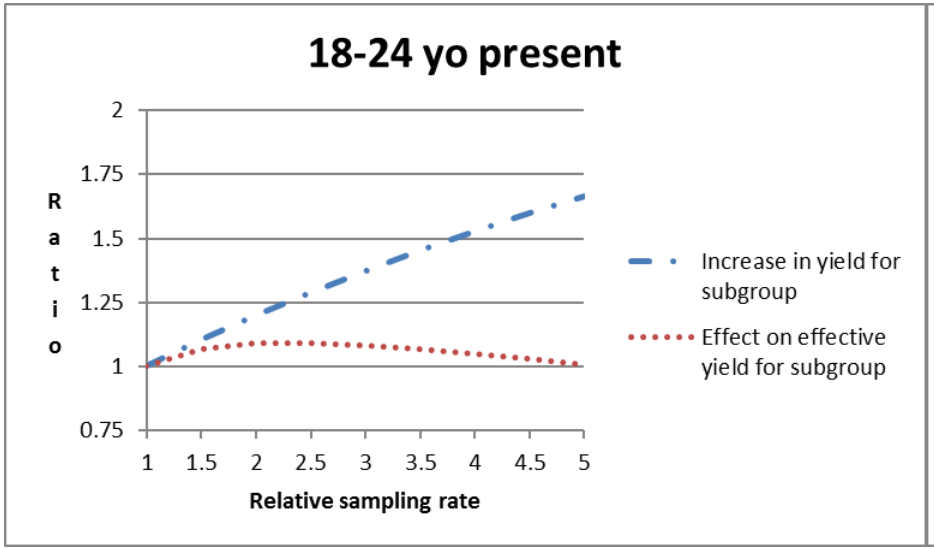
Goal 2: Examine Potential for Oversampling
Good candidates

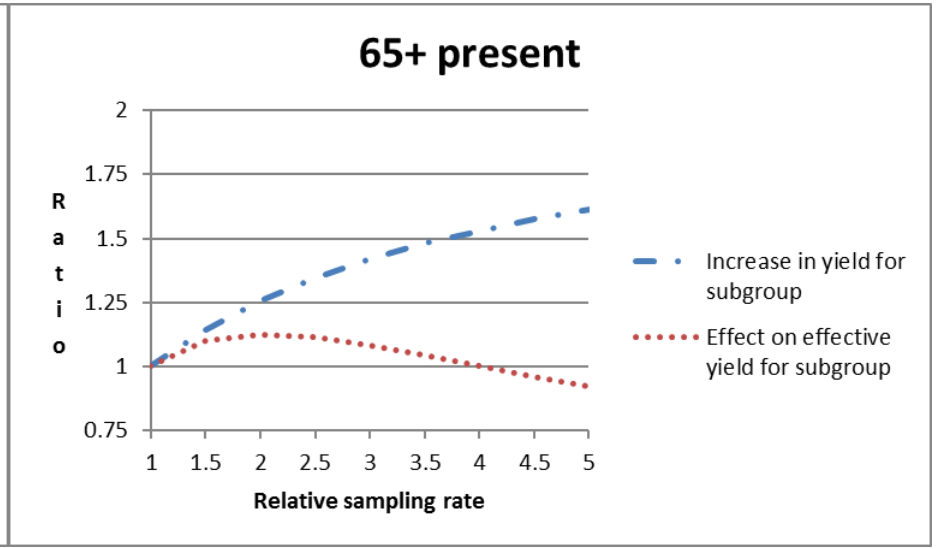
NHTS: Home tenure



Goal 2: Examine Potential for Oversampling
Good candidates

HINTS: Presence of age 18-24 and 65+





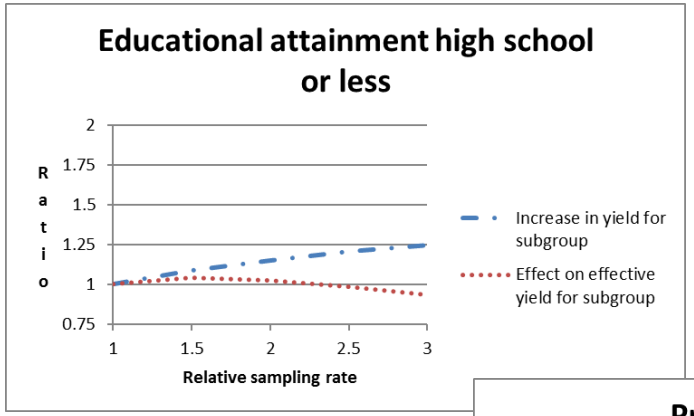
Goal 2: Examine Potential for Oversampling
Good candidates

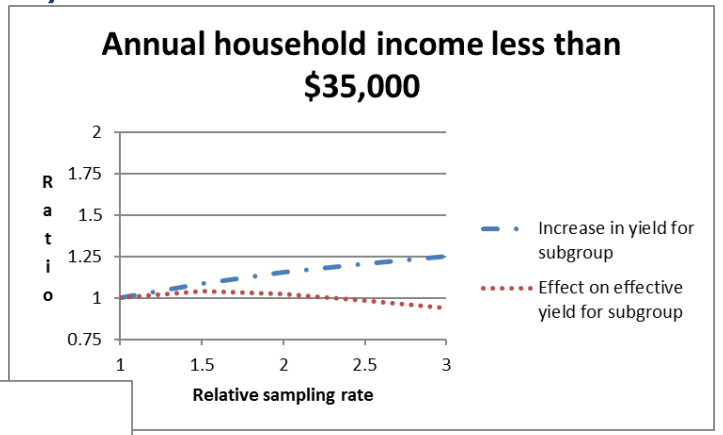
Optimum Oversampling Rates

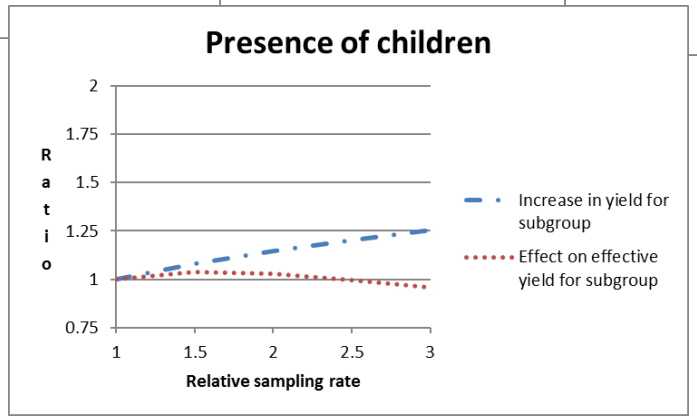
Characteristic	Optimum oversampling rate
NHTS Ethnicity	3.2
NHTS Hispanic surname	2.8
HINTS Ethnicity	3.5
HINTS Hispanic surname	3.7
NHTS Home tenure	1.9
HINTS Presence of age 18-24	2.3
HINTS Presence of age 65+	2.0

Goal 2: Examine Potential for Oversampling

NHTS: Educational Attainment, Income, Presence of children

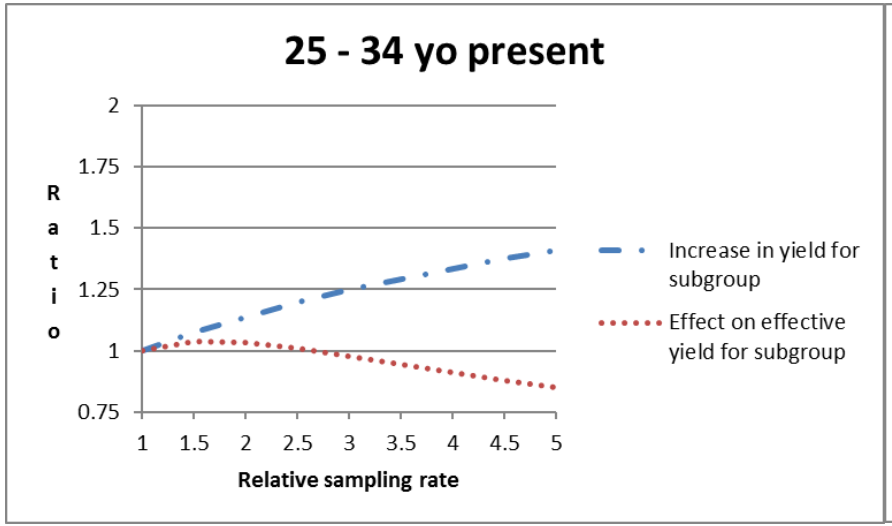


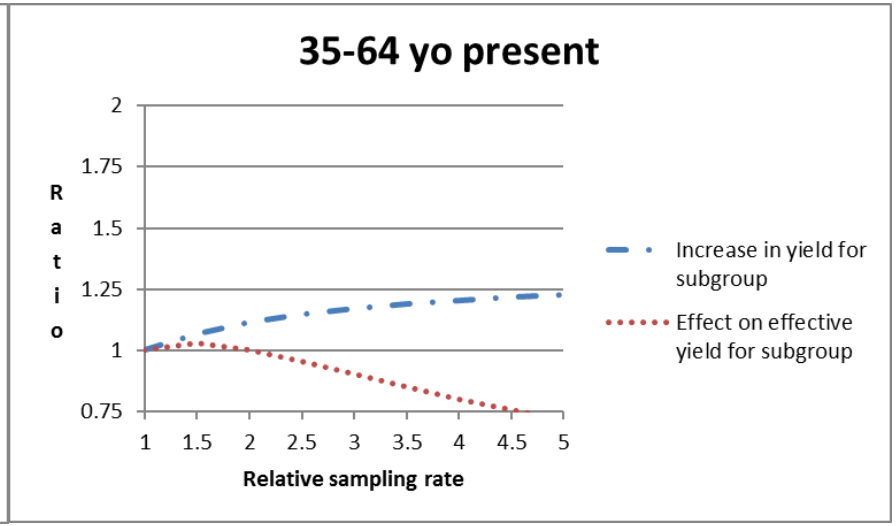




Goal 2: Examine Potential for Oversampling

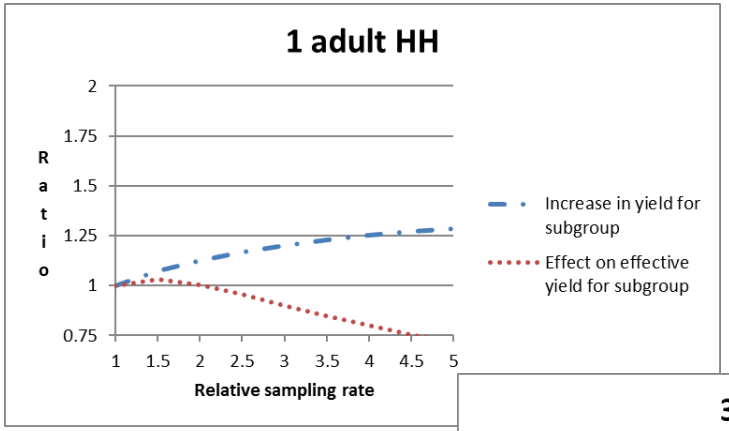
HINTS: Presence of age groups 25-34 and 35-64

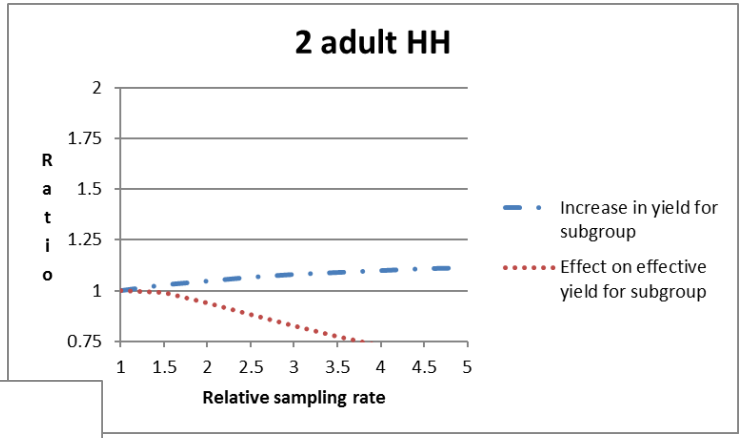


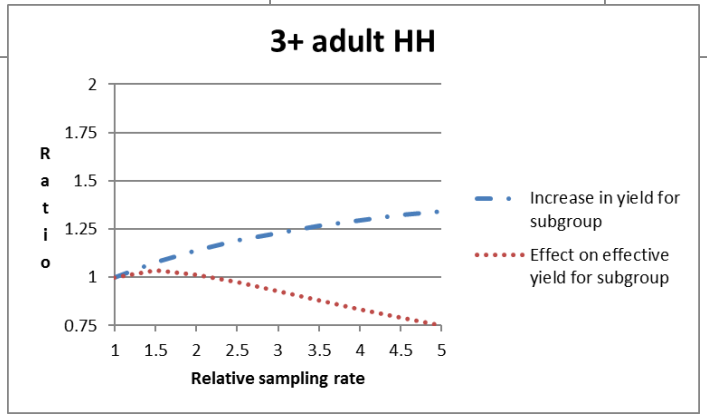


Goal 2: Examine Potential for Oversampling

HINTS: Number of adults

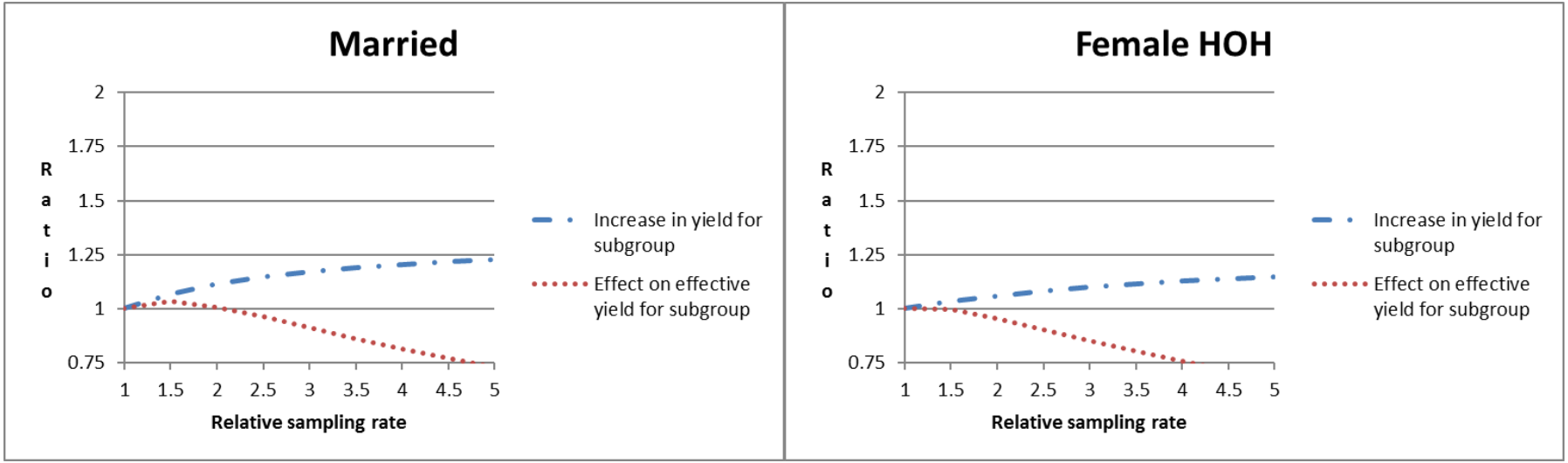






Goal 2: Examine Potential for Oversampling

HINTS: Marital Status, Female Head of Household



Goal 3: Examine Potential for Weighting Adjustments
Preliminary Research

- Classification trees included nonresponse adjustment auxiliary variables used for weighting + appended frame variables
- SAS high performance procedure HPSPLIT used to create trees
- Preliminary findings indicate some potential for using appended variables for nonresponse adjustment, most notably presence of age 65+

Conclusions

Improvements in Appended Data

- Improvements have been made in availability and data quality
  - Lower missingness rates
  - Better agreement between frame variables and survey responses
- Potential variables for oversampling
  - Ethnicity or Hispanic surname
  - Home tenure, presence of age groups 18-24 and 65+

Further Research

- Complete our investigation of potential utility of appended variables for nonresponse adjustment
  - Expand classification tree analyses to identify appended variables which may be related to response propensity
  - Examine associations of appended variables with key survey outcome variables to assess their use in determining potential nonresponse bias
- Repeat the analyses with a different study

Westat

Thank you!

ShelleyBrock@Westat.com
AndrewCaporaso@Westat.com
JillDeMatteis@Westat.com

Address Deliverable Rates

NHTS

ABS Frame Characteristic				Address deliverable rate	
		Total number of addresses	Percent of addresses	Percent	s.e.
Seasonal address	Yes	4,321	0.8	59.4	1.5
	No	924,756	99.2	92.7	0.05
Vacant address	Yes	22,746	2.7	38.1	0.76
	No	906,331	97.3	93.9	0.07

Address Deliverable Rates (cont.)

HINTS

ABS Frame Characteristic				Address deliverable rate	
		Total number of addresses	Percent of addresses	Percent	s.e.
Seasonal address	Yes	70	0.8	73.4	5.7
	No	13,290	99.2	87.9	0.3
Vacant address	Yes	912	6.9	16.3	1.5
	No	12,448	93.1	93.1	0.3

Recruitment/Screener Response Rates

NHTS

Characteristic	Description	Total number of eligible addresses	Percent of eligible addresses	Recruitment response rate	
				Percent	s.e.
Carrier route type	PO Box	9,214	0.8	27.4	1.26
	City delivery	508,817	63.9	29.0	0.09
	Highway contract	20,047	1.9	33.8	0.79
	Rural route	322,356	33.4	33.0	0.17
Dwelling unit type	M: multi-family	198,252	23.9	23.1	0.11
	S: single family	652,968	75.3	32.7	0.08
	P: PO box	9,214	0.8	27.4	1.26
Seasonal address	Yes	2,480	0.5	37.9	1.83
	No	857,954	99.5	30.3	0.07
Vacant address	Yes	8,669	1.1	17.1	0.73
	No	851,765	98.9	30.5	0.08
Drop point address	Yes	9,572	1.6	23.3	0.66
	No	850,862	98.4	30.5	0.07
PO box only way to get mail	Yes	9,214	0.8	27.4	1.26
	No	851,220	99.2	30.4	0.07
Telephone match	Yes	276,696	33.4	38.2	0.12
	No	583,738	66.6	26.5	0.09
Surname available	Yes	774,271	90.4	31.7	0.08
	No	86,163	9.6	18.4	0.26

Recruitment/Screener Response Rates (cont.)

HINTS

Characteristic	Description	Total number of eligible addresses	Percent of eligible addresses	Response rate	
				Percent	s.e.
Carrier route type	PO Box	795	7.3	31.4	2.2
	City delivery	7,964	58.7	31.9	0.6
	Highway contract	152	1.9	35.5	4.2
	Rural route	2,857	32.1	35.0	1.0
Dwelling unit type*	M: multi-family	3,102	21.8	24.1	0.9
	S: single family	7,871	70.9	35.8	0.7
	P: PO box	795	7.3	31.4	2.2
Seasonal address	Yes	55	0.6	47.0	7.8
	No	11,713	99.4	32.8	0.5
Vacant address	Yes	184	1.3	19.8	4.4
	No	11,584	98.7	33.1	0.5
Drop point address	Yes	223	1.5	34.1	4.7
	No	11,545	98.5	32.9	0.5
Telephone match	Yes	3,193	30.3	43.0	0.8
	No	8,575	69.7	28.6	0.6
Surname available	Yes	10,031	86.7	34.6	0.6
	No	1,737	13.3	21.7	1.0
Age group available	Yes	9,919	85.7	34.6	0.6
	No	1,849	14.3	22.7	1.0
Number of HH Adults available	Yes	9,919	85.7	34.6	0.6
	No	1,849	14.3	22.7	1.0

Missing Rates for Appended Demographics

NHTS

Characteristic	Entire NHTS sample			NHTS recruitment respondents only			NHTS retrieval respondents only		
	N	Percent missing	s.e.	n	Percent missing	s.e.	n	Percent missing	s.e.
TOTAL	929,077			252,304			129,696		
Education	256,550	27.4	0.05	52,973	21.1	0.16	27,611	23.8	0.32
Ethnicity	256,550	27.4	0.05	52,973	21.1	0.16	27,611	23.8	0.32
Gender of HoH	123,490	12.7	0.06	17,251	6.6	0.09	8,661	8.9	0.16
Household income	123,490	12.7	0.06	17,251	6.6	0.09	8,661	8.9	0.16
Marital status	245,946	26.0	0.07	54,856	21.6	0.13	28,902	22.7	0.17
Home tenure	191,459	20.0	0.09	32,726	12.5	0.17	16,628	16.3	0.30

Missing Rates for Appended Demographics (cont.)

NHES

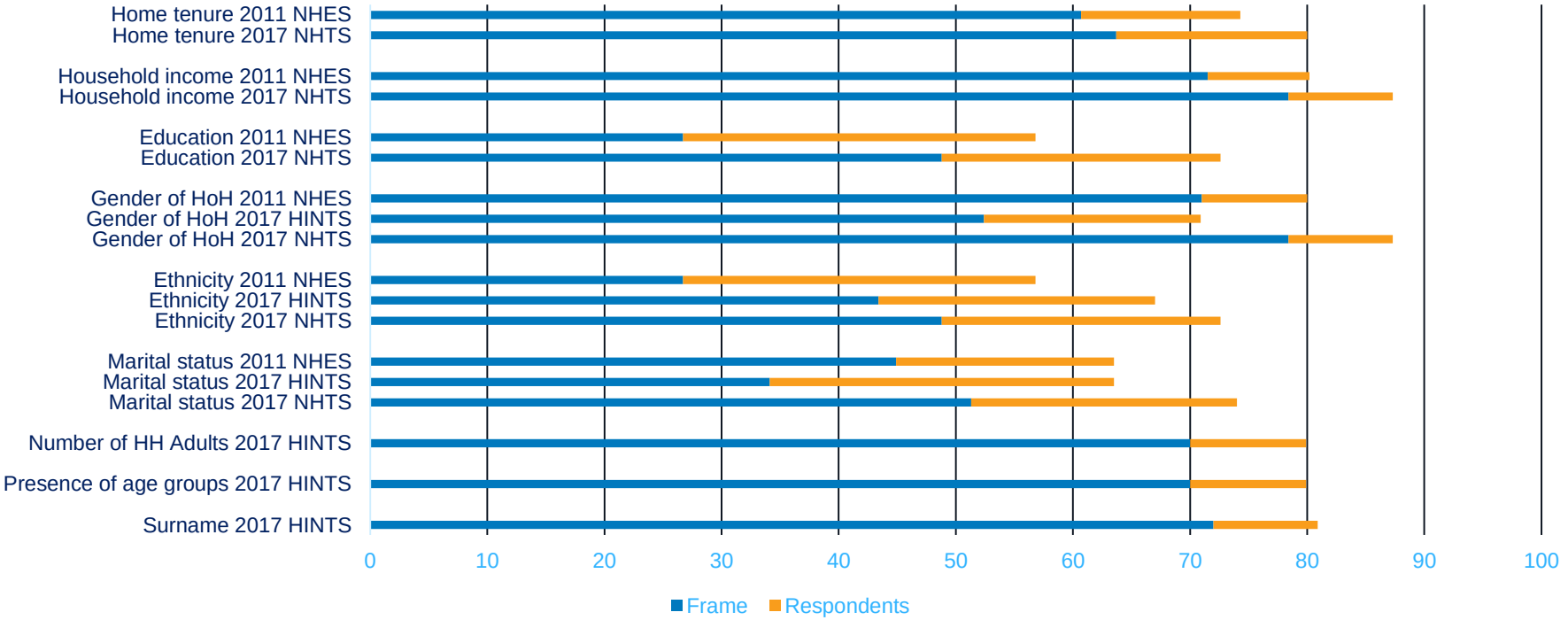
Characteristic	Entire NHES sample			NHES respondents only		
	N	Percent Missing	s.e.	n	Percent Missing	s.e.
TOTAL	41,264			5,587		
Education	17,945	43.2	0.21	1,682	30.1	0.74
Ethnicity	17,946	43.2	0.21	1,682	30.1	0.74
Gender	8,415	20.0	0.17	501	9.0	0.49
Household income	8,324	19.8	0.17	491	8.7	0.48
Marital status	15,201	36.5	0.21	1079	18.6	0.58
Home tenure	10,760	25.7	0.21	781	13.6	0.56

Missing Rates for Appended Demographics (cont.)

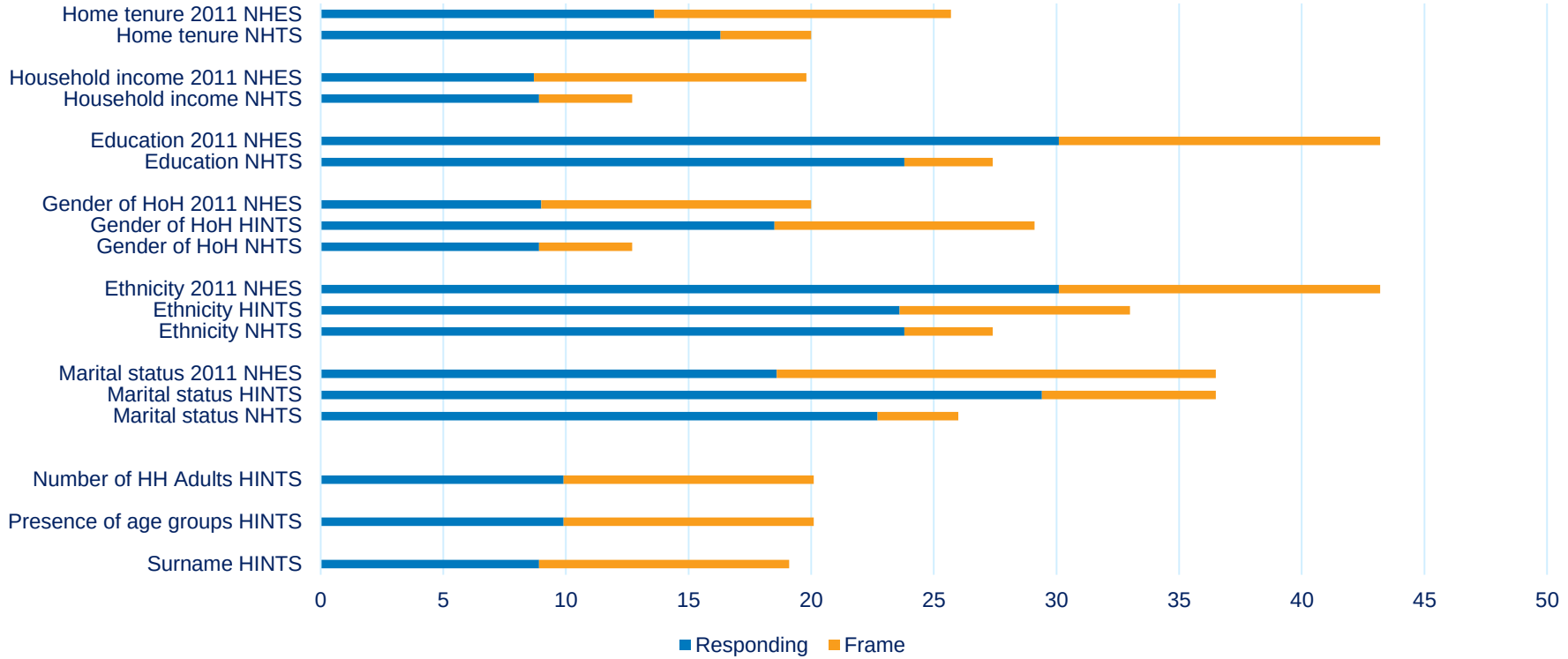
HINTS

Characteristic	Entire HINTS sample			HINTS eligible households			HINTS respondents only		
	N	Percent Missing	s.e.	n	Percent Missing	s.e.	n	Percent Missing	s.e.
TOTAL	13,360			11,768			3,335		
Ethnicity	4,467	33.0	0.53	3,384	27.9	0.55	811	23.6	0.79
Gender of HoH	3,994	29.1	0.47	2,945	23.7	0.47	649	18.5	0.70
Marital status	4,847	36.5	0.55	3,750	31.6	0.57	977	29.4	1.00
Number of HH Adults	2,787	20.1	0.47	1,849	14.3	0.41	356	9.9	0.52
Presence of age groups	2,787	20.1	0.47	1,849	14.3	0.41	356	9.9	0.52

Non-missing Rates for Appended Demographics



Missingness Rates for Appended Demographics



Potential for Weighting Adjustments

HINTS

- One tree created with the following appended variables predicting response
  - Presence of age 65+
  - Hispanic surname
  - 1 adult household
  - Gender of head of household

Potential for Weighting Adjustments

NHTS

Geographic Area	Presence of children	Hispanic	Hispanic surname	Home tenure	Educational attainment	Household income	Presence of 18-24 yo	Presence of 25-34 yo	Presence of 35-64 yo	Presence of 65+ yo
Arizona		x		x					x	x
California		x	x	x	x				x	x
Des Moines						x			x	x
Georgia		x		x	x				x	x
Tulsa (Census blocks)									x	x
Northern Iowa				x					x	x
Maryland				x	x				x	x
North Carolina		x		x	x				x	x
North Central Texas (counties)		x		x	x				x	x
New York		x		x	x				x	x
South Carolina		x		x	x	x			x	x
Texas		x	x	x	x	x			x	x
Wisconsin	x	x		x	x	x			x	x
Census Division 1									x	x
Census Division 2, excluding NY										x
Census Division 3, excluding WI		x		x	x				x	x
Census Division 4, excluding IA blocks				x					x	x
Census Division 5, excluding GA, MD, NC, SC		x		x	x				x	x
Census Division 6				x					x	x
Census Division 7, excluding TX and OK blocks										x
Census Division 8, excluding AZ		x						x	x	x
Census Division 9, excluding CA				x	x				x	x