Does Undercoverage on the US Address-based Sampling Frame Translate to Coverage Bias?

Ashley Amaya, Stephanie Zimmer, Katherine Morton, and Rachel Harter

July 20, 2017

The ABS Frame

- US Postal System's Computerized Delivery Sequence File (CDS)
  - Contains all addresses for which USPS delivers mail
    - 90–98% estimated coverage of residential housing units (AAPOR 2016)
    - Most addresses use the format:

      123 Main Street

      Unit 1

      Anytown, NY 12345
    - Names are not included

Undercoverage

- Undercoverage is much higher in rural areas
  - 23-35% in rural areas vs. 1-10% in urban areas (Dohrmann et al 2006; Dohrmann et al 2007; O'Muircheartaigh et al 2007)

- The CDS frame
  - Purposely excludes:
    - Unique ZIP codes (e.g., Indian reservations and universities)
    - Vacant units in rural areas

  - Includes "unusable" addresses:
    - PO Boxes
    - Simplified addresses

Coverage Bias

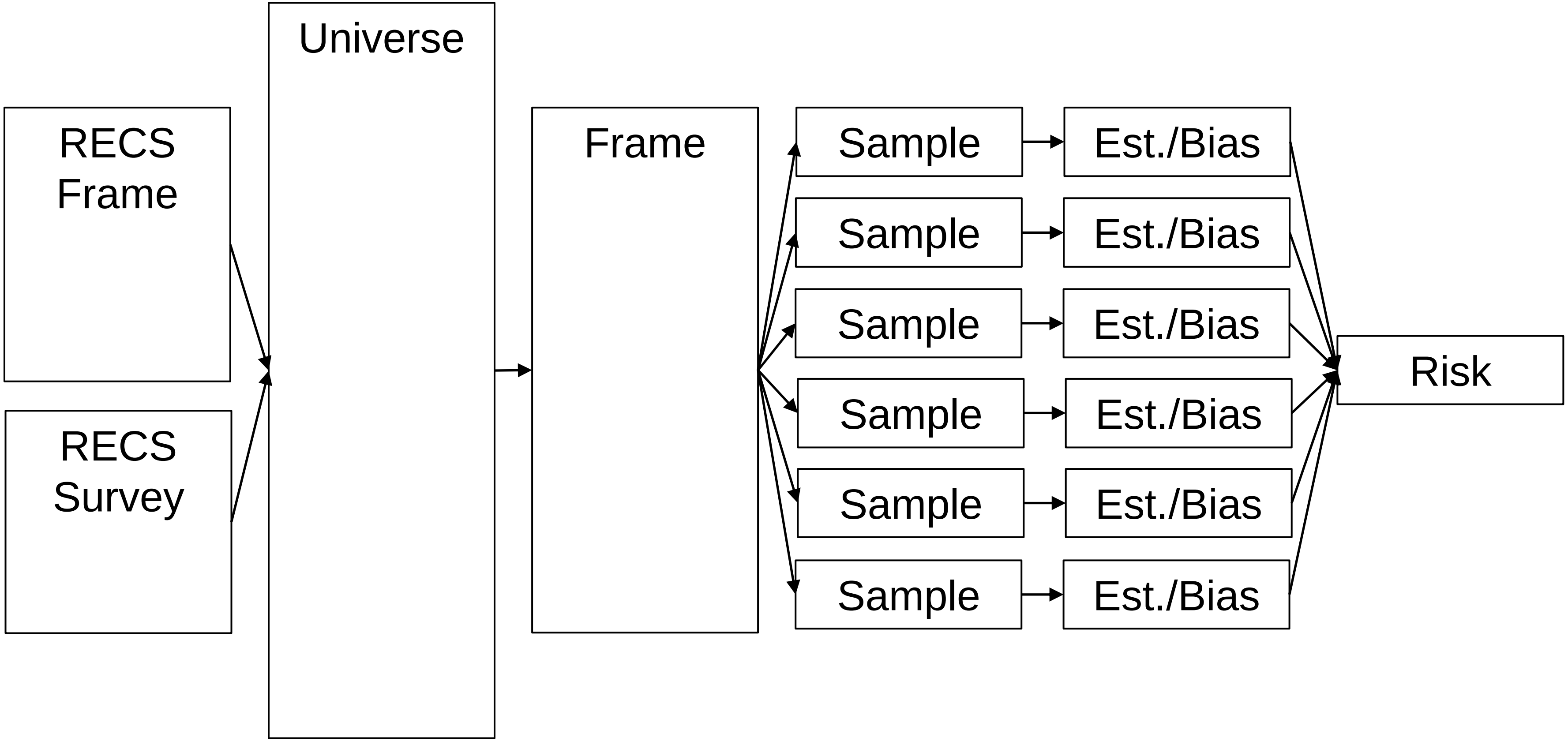
- Three studies have assessed the impact of undercoverage on bias
  - English et al (2011)
    - Fertility in Cumberland, Maine
  - Morton et al (2010)
    - Substance abuse with small uncovered counts
  - Eckman & Kreuter (2013)
    - Fertility, health, sexuality, and demographics of two list frames (not the CDS)

This Presentation

- Research Question
  - What is the risk of coverage bias when using the USPS CDS in a face-to-face survey?

- Goal
  - Inform decisions on whether to
    - Use the ABS frame for a given survey, and/or
    - Enhance the ABS frame (e.g., a hybrid design or HOI)

Methods – Monte Carlo Simulation



Methods – Monte Carlo Simulation

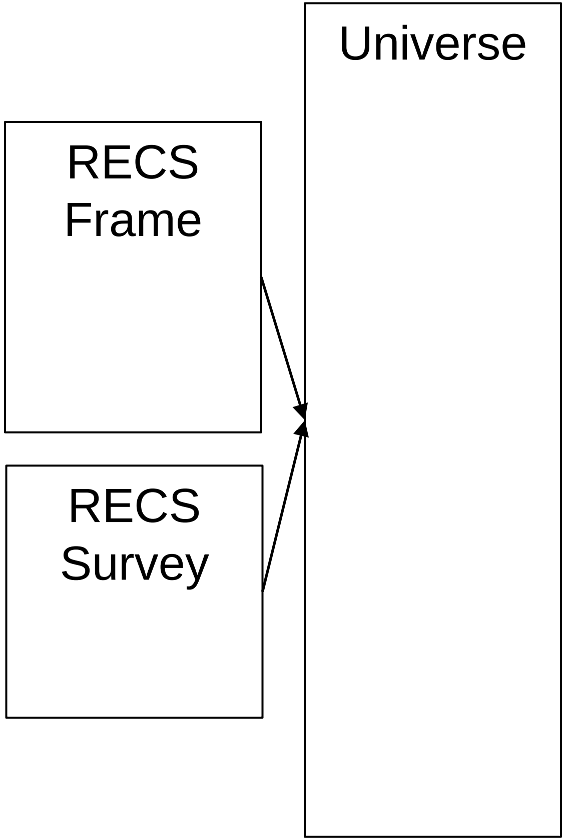
RECS Frame

- 800 Census block groups across the US
- 579,459 CDS addresses
- 6,841 enumerated addresses

RECS Survey

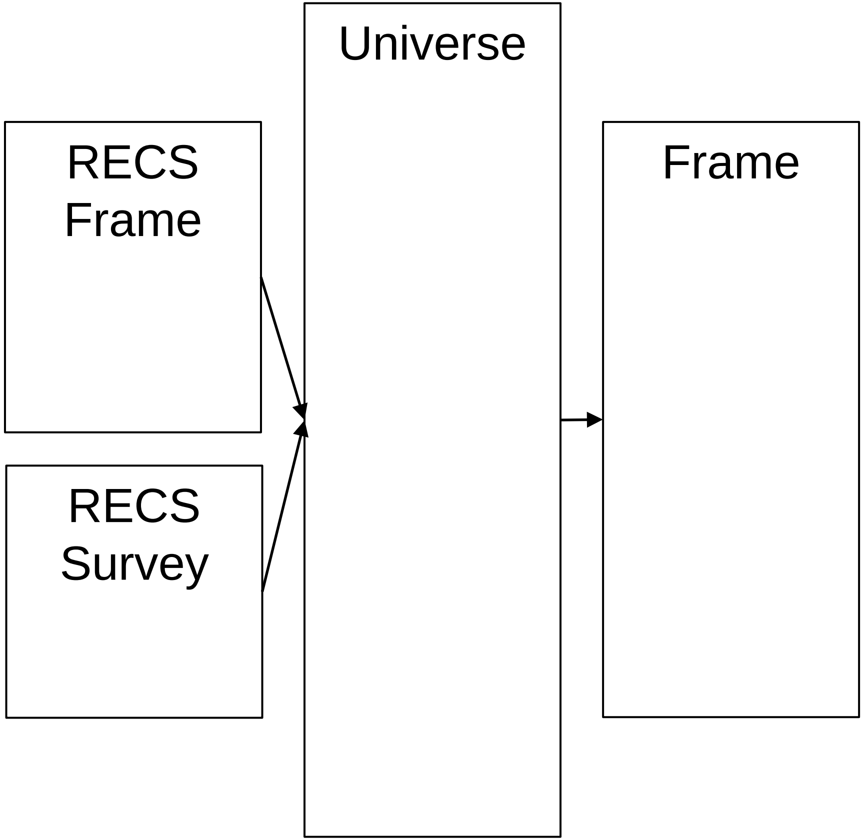
- 12 demographic and building characteristic variables

Methods – Monte Carlo Simulation



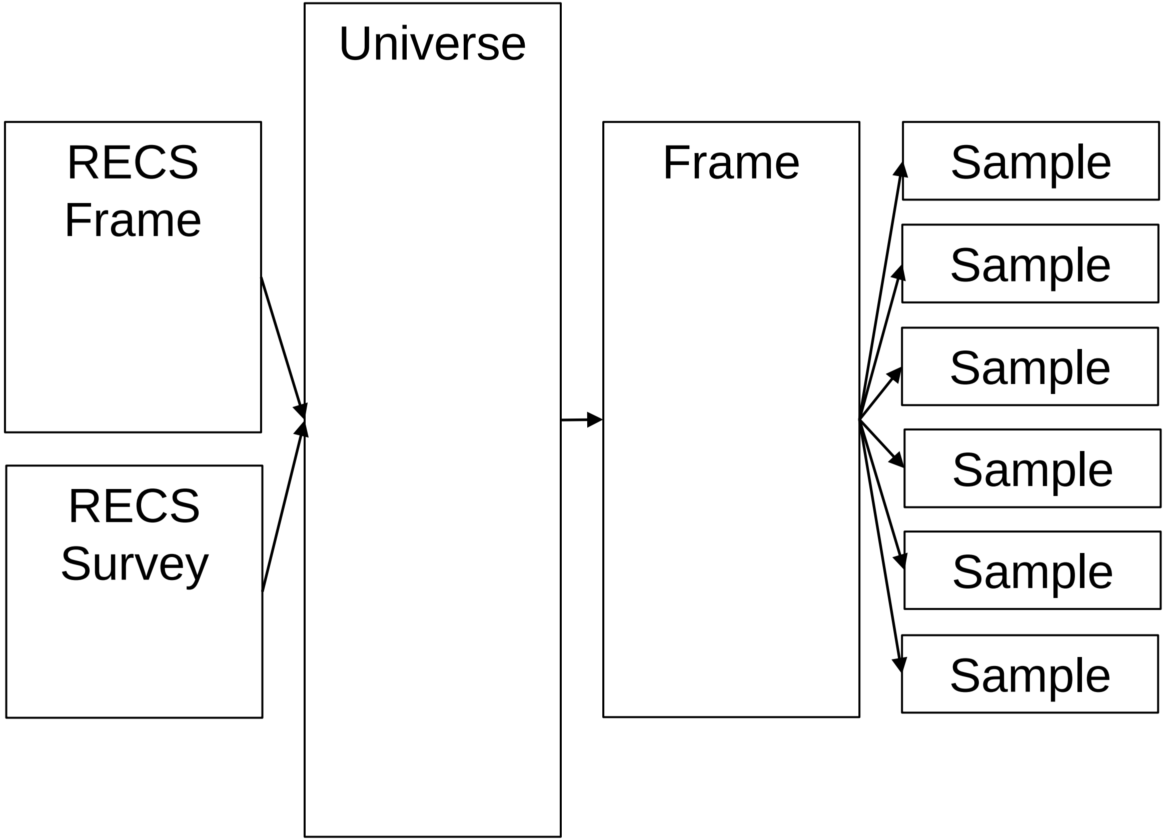
- Created one universe
- Replicated cases from the RECS survey by their final weights
- Used the frame information (and appended ACS data) to assign coverage propensities

Methods – Monte Carlo Simulation



- Created one frame for each coverage rate 1-100% (n=100)
- Assigned each unit a coverage propensity

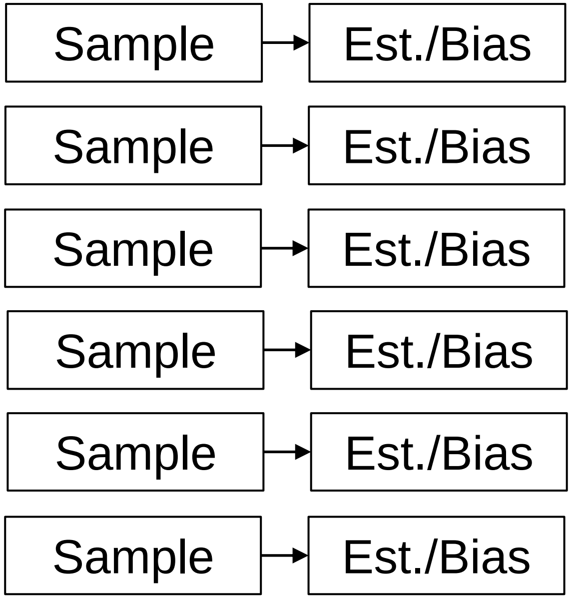
Methods – Monte Carlo Simulation



- For each frame
  - Drew 1,000 samples of 1,000 addresses per sample
- 2-stage design
  - 200 PSUs
  - 5 addresses per PSU

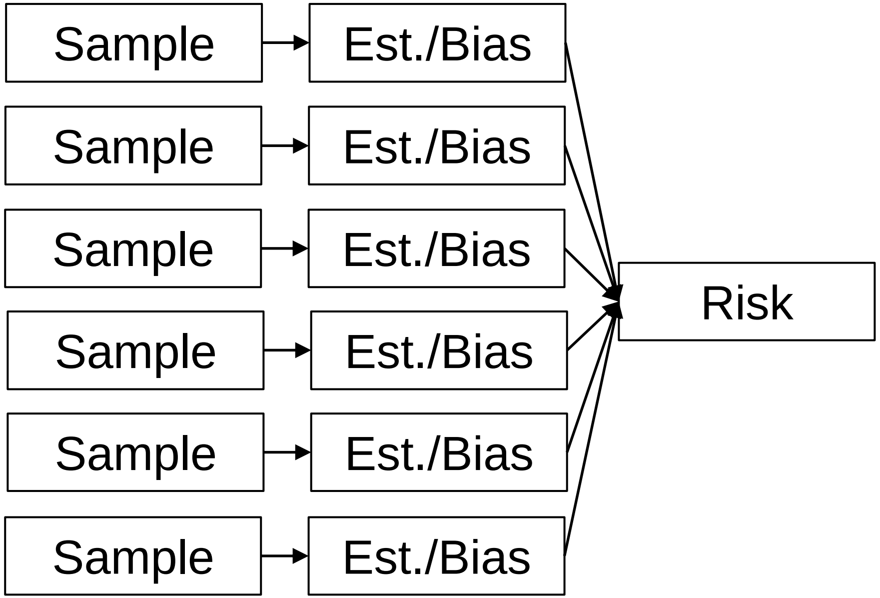
Methods – Monte Carlo Simulation

- For each sample,
  - Calculated the proportion/distribution/mean of each of the 12 variables
  - Calculated bias compared to the universe
    - Bias ($\hat{\theta} - \theta$) and relative bias ($\frac{\hat{\theta}-\theta}{\theta}$)
    - Z-test for significance

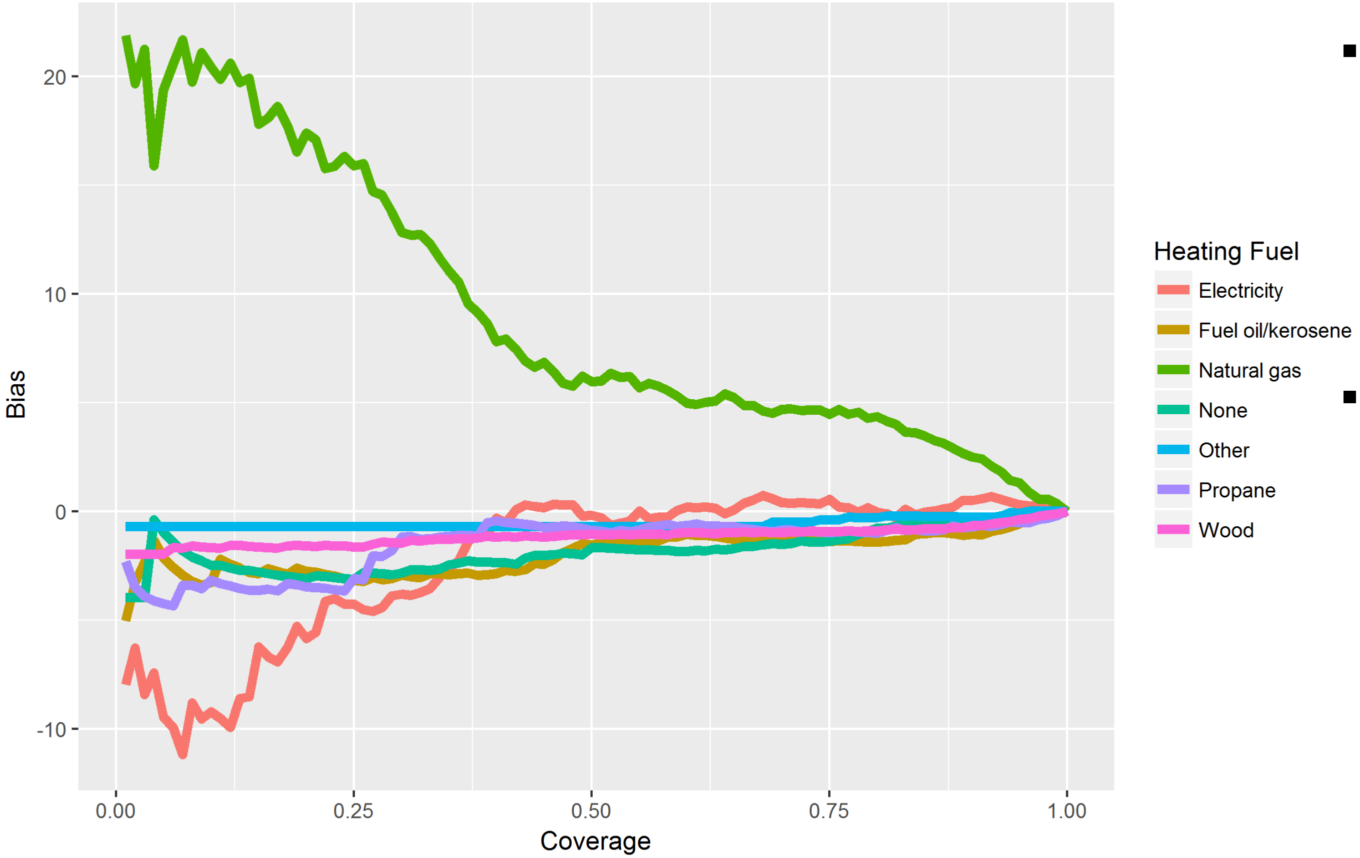


Methods – Monte Carlo Simulation

- For each level of coverage,
  - Risk is the proportion of samples for which the estimate was significantly different than the universe ($p<0.05$)

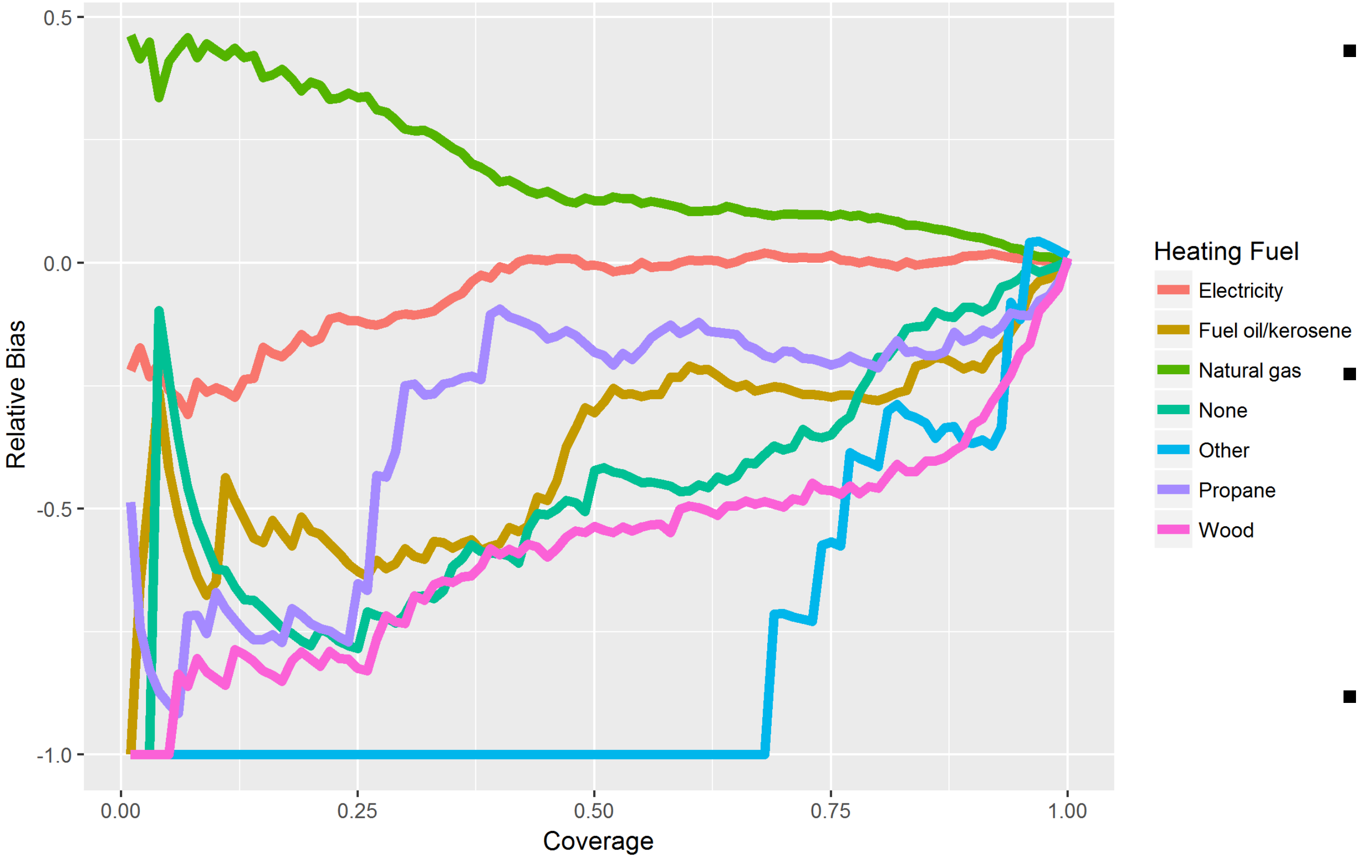


Modeled Coverage Distribution – Heating Fuel Bias



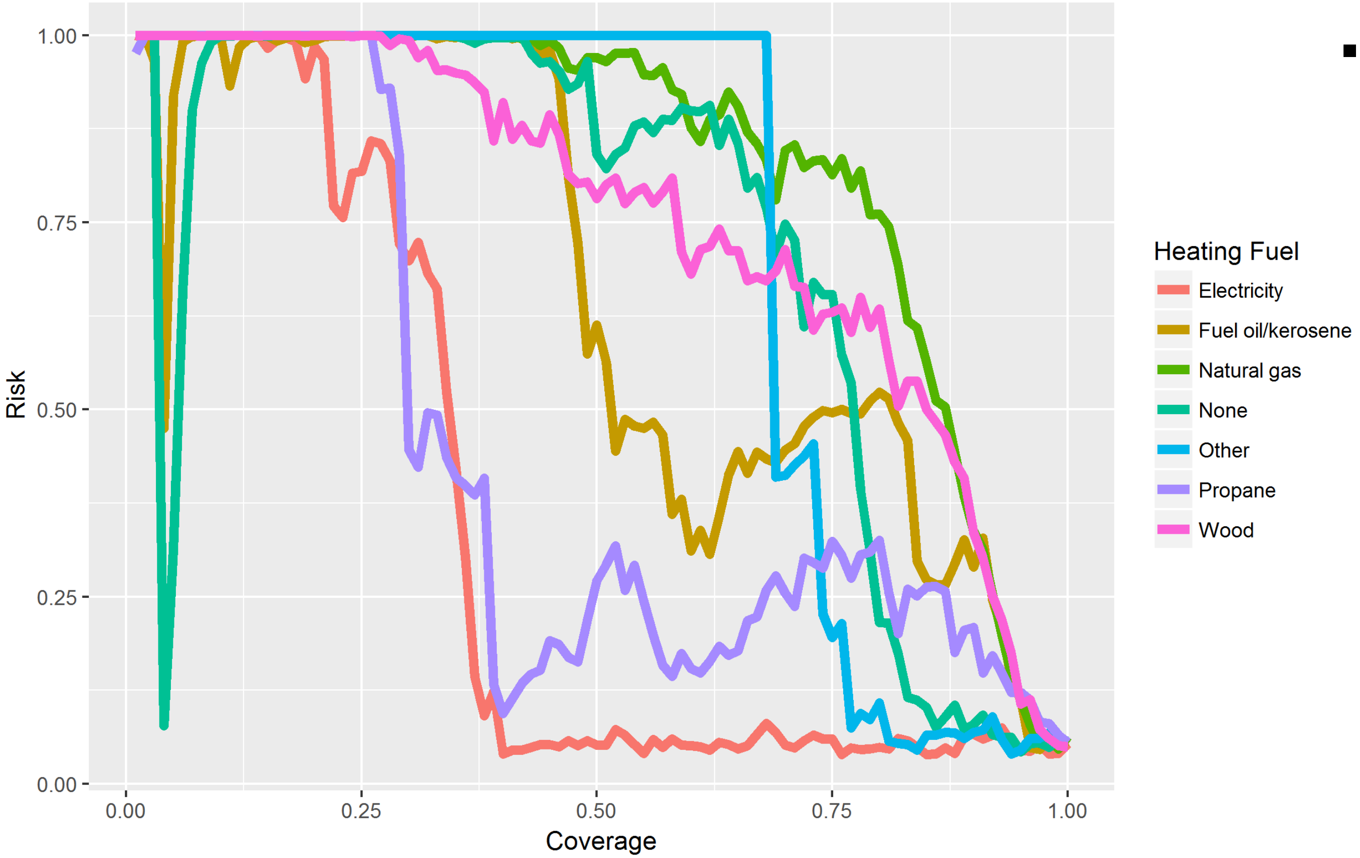
- As coverage declines, quickly begin to overest. natural gas heating
- Other heating fuels are relatively stable until coverage drops below ~50%.

Modeled Coverage Distribution – Heating Fuel Relative Bias



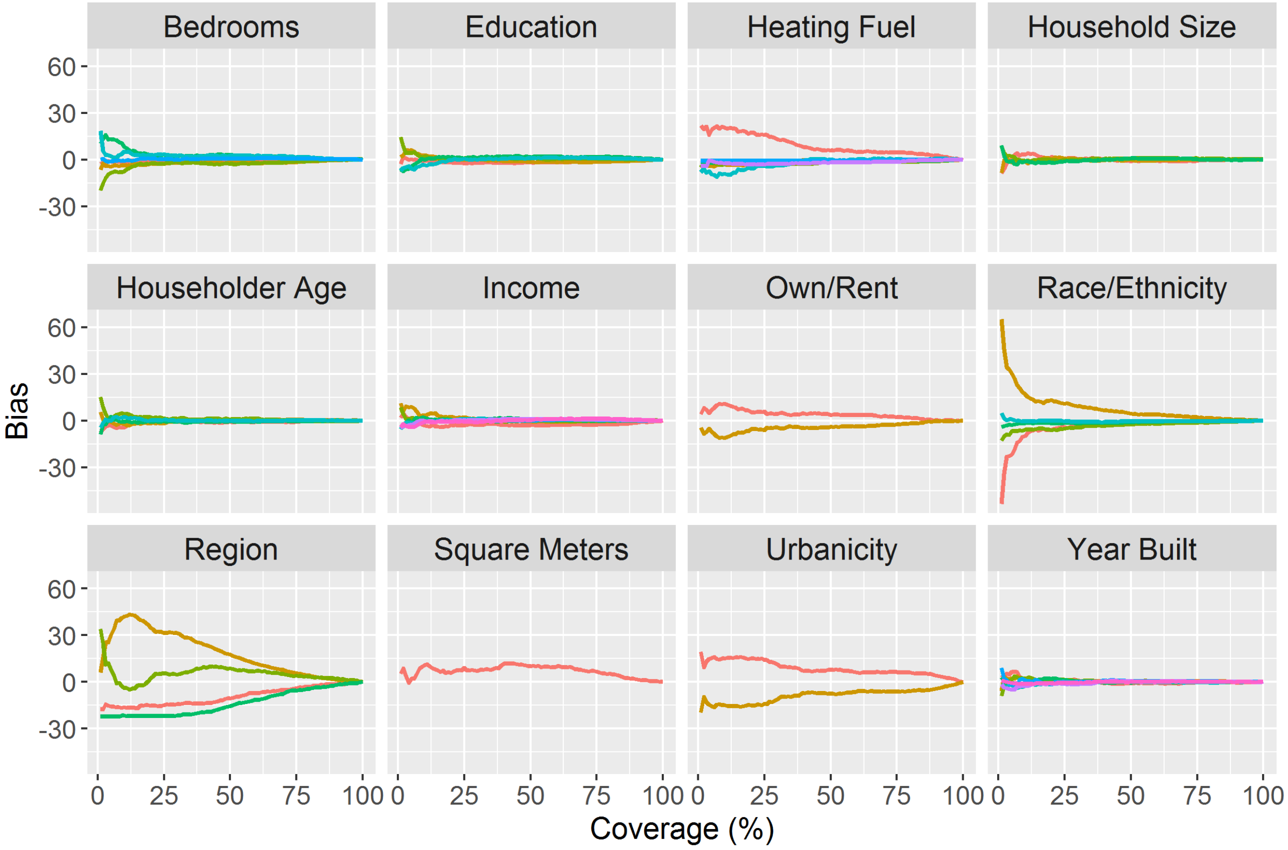
- The magnitude is small but meaningful since prevalence is small.
- Relative bias increases quickly (except elec.) as coverage declines.
- Findings not surprising. Coverage & heating fuel both corr. with urbanicity.

Modeled Coverage Distribution – Heating Fuel Risk



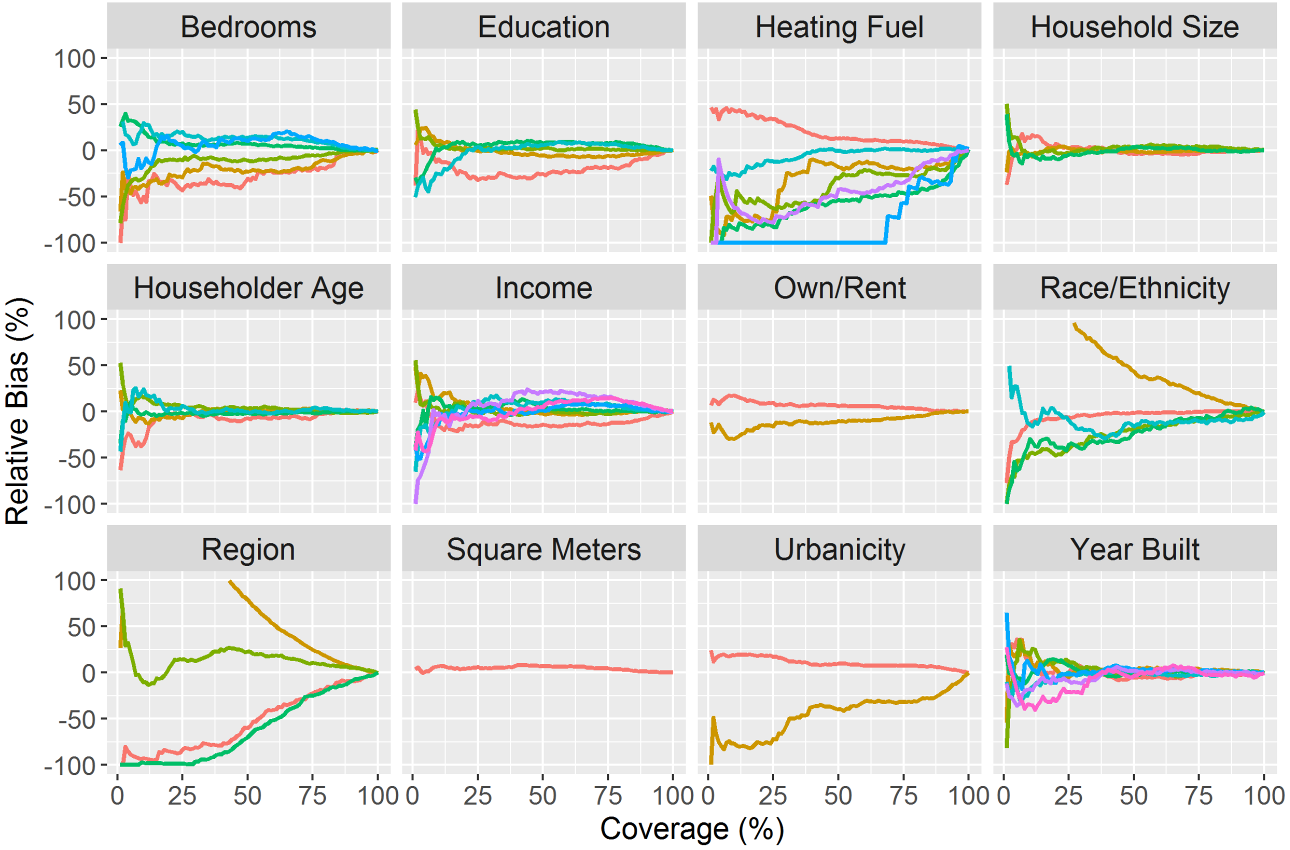
- Risk increases quickly for most heating fuels as coverage declines.

Modeled Coverage Distribution - Bias



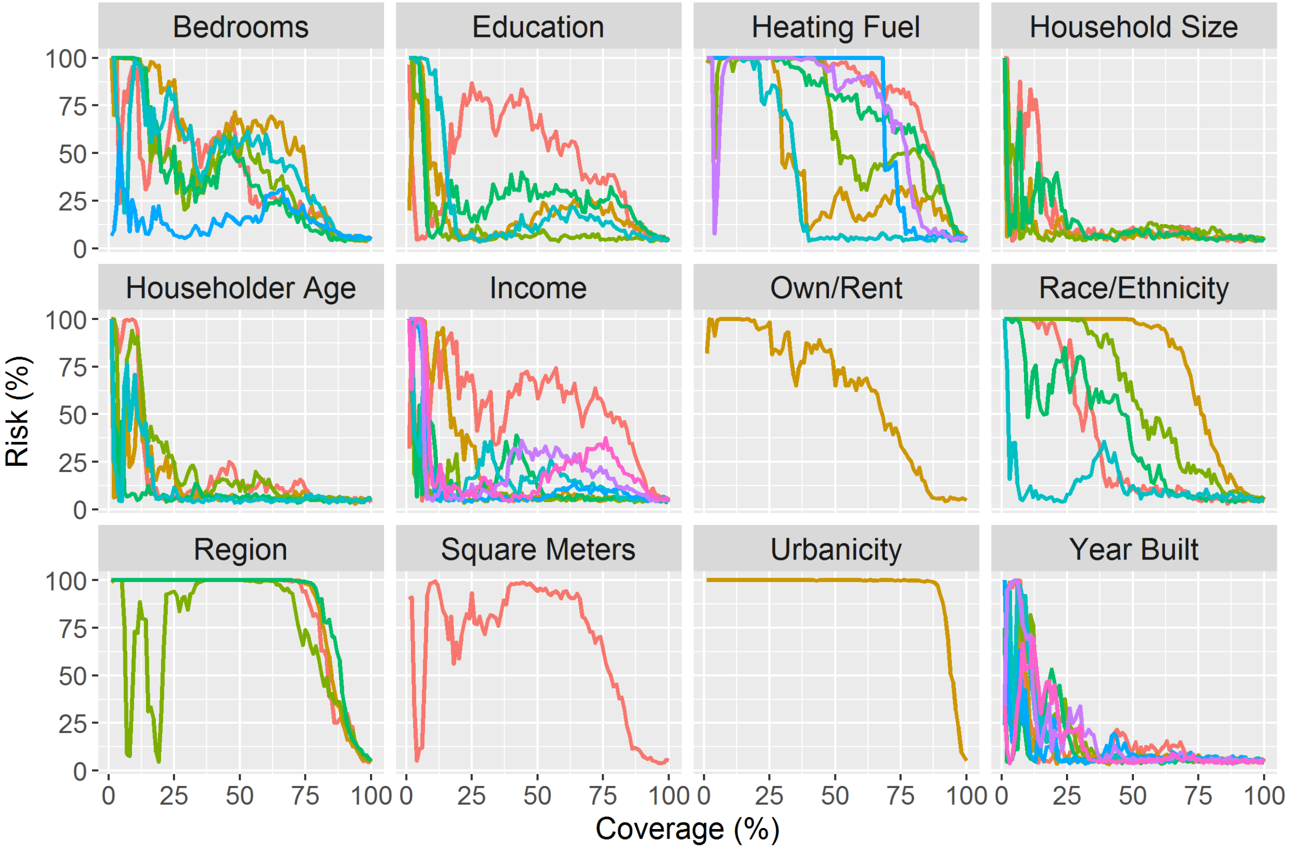
- The magnitude of the bias is relatively unaffected by coverage for 50% of the variables.

Modeled Coverage Distribution – Relative Bias



- Only 25% variables are relatively unaffected by coverage when considering relative bias.
- Bedrooms and education had small changes, but had large effect given small prevalence.

Modeled Coverage Distribution – Risk



- Risk is dependent on the variable of interest.
  - HH size unaffected.
  - Year built and age has low risk when coverage > 75%
  - Risk of bias increases quickly for other variables as coverage declines.

Summary

- What is the risk of coverage bias when using the USPS CDS in a face-to-face survey?
  - It depends on:
    - The variable of interest
    - The unit of analysis (categorical or dichotomous)
    - The level of coverage

Next Steps

- Replicate
  - Simulate other sub-national domains: Rural and Mid-Atlantic
  - Recreate analysis for alternative modes: Mail
  - RECS frame may not be the true universe
    - Did not attempt to enhance CDS in high coverage areas
  - RECS is not necessarily applicable to a wide variety of surveys (e.g., health)
- Determine whether weights could reduce risk
- Identify patterns in bias risk by variable type

Citations

- American Association for Public Opinion Research. (2016). *Address-based sampling* (prepared for AAPOR Council by the Task Force on Address-Based Sampling; R. Harter, Chair). Oakbrook Terrace, IL: Author. Retrieved from http://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx
- Dohrmann, S., Han, D., & Mohadjer, L. (2006). Residential address lists vs. traditional listing: Enumerating households and group quarters. In *Proceedings of the 2006 Joint Statistical Meetings, American Statistical Association, Survey Research Methods Section, Seattle, WA* (pp. 2959-2964). Alexandria, VA: American Statistical Association.
- Dohrmann, S., Han, D., & Mohadjer, L. (2007). Improving coverage of residential address lists in multistage area samples. In *Proceedings of the 2007 Joint Statistical Meetings, American Statistical Association, Section on Survey Research Methods, Salt Lake City, UT* (pp. 3219-3126). Alexandria, VA: American Statistical Association.
- Eckman, S., & Kreuter, F. (2013). Undercoverage rates and undercoverage bias in traditional housing unit listing. *Sociological Methods and Research, 42*, 264-293. https://doi.org/10.1177/0049124113500477
- English, N., Dekker, K., & O'Muircheartaigh, C. (2011). Choices of Frame Construction on the National Children's Study: Impacts on Address Quality and Survey Results. In *Proceedings of the 2011 Joint Statistical Meetings, American Statistical Association, Section on Survey Research Methods, Miami Beach, FL* (pp. 4250-4259). Alexandria, VA: American Statistical Association.
- Morton, K., McMichael, J., Ridenhour, J., & Bose, J. (2010). Address-based sampling and the National Survey on Drug Use and Health: Evaluating the effects of coverage bias. In *Proceedings of the 2010 Joint Statistical Meetings, American Statistical Association, Section on Survey Research Methods, Vancouver, British Columbia* (pp. 4902-4907). Alexandria, VA: American Statistical Association.
- O'Muircheartaigh, C., English, N., & Eckman, S. (2007). Predicting the relative quality of alternative sampling frames. In *Proceedings of the 2007 Joint Statistical Meetings, American Statistical Association, Section on Survey Research Methods, Salt Lake City, UT* (pp. 3239-3248). Alexandria, VA: American Statistical Association.

More Information

Ashley Amaya

Research Survey Methodologist

+1.202.728.2486

aamaya@rti.org

Coverage Propensity Model (RECS Frame, n=586,301)

Variable	Beta
Intercept	-1.13***
Urbanicity (ref=rural)	
Urban Cluster	-0.11**
Urban Area	1.09***
Building Type (ref=multi-family unit)	
Single Family Unit	2.24***
Unknown	-3.83***
Region (ref=West)	
Northeast	-1.33***
Midwest	2.07***
South	0.45***
Mean Income in CBG (in $1,000s)	0.05***
CBG Race/Ethnicity	
Percent Hispanic	0.06
Percent NH Black	4.95***
Percent NH Oth	-0.17
CBG Education	
High School Graduate	5.38***
Bachelors Degree +	1.27***
Percent Home Owners in CBG	0.26**
Percent Vacant HUs in CBG	-7.76***