Small domain estimation using probability and non-probability survey data

Adrijo Chakraborty and N Ganesh

March 2018, FCSM

NORC at the UNIVERSITY of CHICAGO

Outline

- Introduction to the Problem
- Approaches for Combining Probability and Non-Probability Samples
- Investigation of two Small Area Models
- Real data analysis
- Summary

Introduction

- For cost reasons, some studies use a combination of probability and non-probability samples
  - Cost associated with obtaining a larger probability sample and/or
  - Cost associated with obtaining sufficient sample size for low incidence target populations

- Given the unknown biases associated with a non-probability sample, what method(s) are best for combining a probability sample with a non-probability sample
  - We want more reliable estimates (hence we use the non-probability sample)
  - But we don’t want to introduce “too much” bias

Combining Probability and Non-Probability Samples

- Different methods to combining
  - Propensity based pseudo weighting methods (Elliott)
  - Model-based methods (Elliott & Valliant, Wang et. al.)
  - Raking / calibration approaches (Fahimi et. al., DiSogra et. al.)
- We investigated approaches that use small area models (Elliott & Haviland)
  - Assume that the probability sample generates unbiased estimates
  - Assume that the non-probability sample estimates are biased
  - Considered two small area models
    1. Model probability sample estimates with non-probability sample estimates as covariates
    2. Bivariate model for probability and non-probability sample estimates

Model 1: Fay-Herriot Model (probability survey data)

- Domains are constructed using race, age, education, gender
- Direct estimates y_d^P from probability sample for domain *d* are unbiased

$$y_d^P = \alpha_d + x_d'\gamma + v_d + e_d^P$$

  - Fixed effect α_d is parametrized by main effects for race, age, gender, education
  - x_d is a vector of domain-level covariates which includes the non-probability sample estimate
  - v_d is a domain-level random effect
  - e_d^P are the sampling errors
- Model-based estimates for domains are derived using a standard prediction approach
- National-level estimates are obtained by aggregating (by population size) the model-based domain-level estimates

Non-probability sample for domain

- Possible bias in non probability survey estimates.
- Extend Fay-Herriot model for non probability survey data.
- We propose additive bias term for each domain.
- Variance estimation using non probability survey data (assuming known domain level variances)

Model 2: Bi-Variate Fay-Herriot Model

- Direct estimates y_d^P from probability sample for domain *d* are unbiased

$$y_d^P = \alpha_d + x'_d\gamma + v_d + e_d^P$$

- Direct estimates y_d^{NP} from non-probability sample for domain *d* are biased

$$y_d^{NP} = \alpha_d + \beta_d + x'_d\gamma + v_d + e_d^{NP}$$

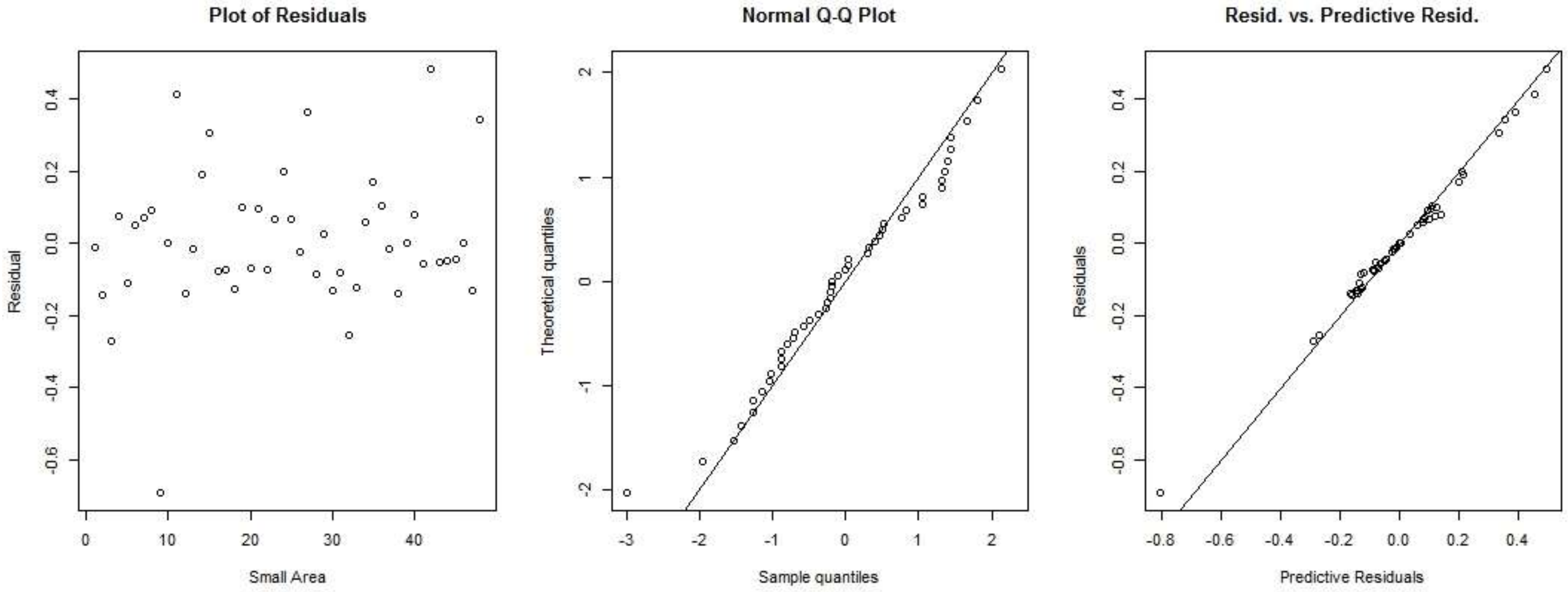
  - Fixed effect α_d is parametrized by main effects for race, age, gender, education
  - **Bias term $\boldsymbol{\beta_d}$** is parametrized by main effects for race, age, gender, education
  - x_d is a vector of domain-level covariates
  - v_d is a domain-level random effect
  - e_d^P and e_d^{NP} are sampling/non-sampling errors
- National-level estimates are obtained by aggregating (by population size) the model-based domain-level estimates

NORC at the UNIVERSITY of CHICAGO

Data Application: Food Allergy Study of 18+ Adults

- ~7,200 probability sample completes
  - Probability sample selected
  - ~33,300 non-probability sample completes
  - Non-probability sample obtained from other sample vendors
- Analyzed 5 measures:
  - Ever had a food allergy
  - Peanut allergy
  - Milk allergy
  - Either biological parent has a food allergy
  - Either biological parent has an environmental allergy
- Constructed 48 domains: Race by Age by Education by Gender

Model 1: Residual Plots (when modeling “Ever had a Food Allergy”)

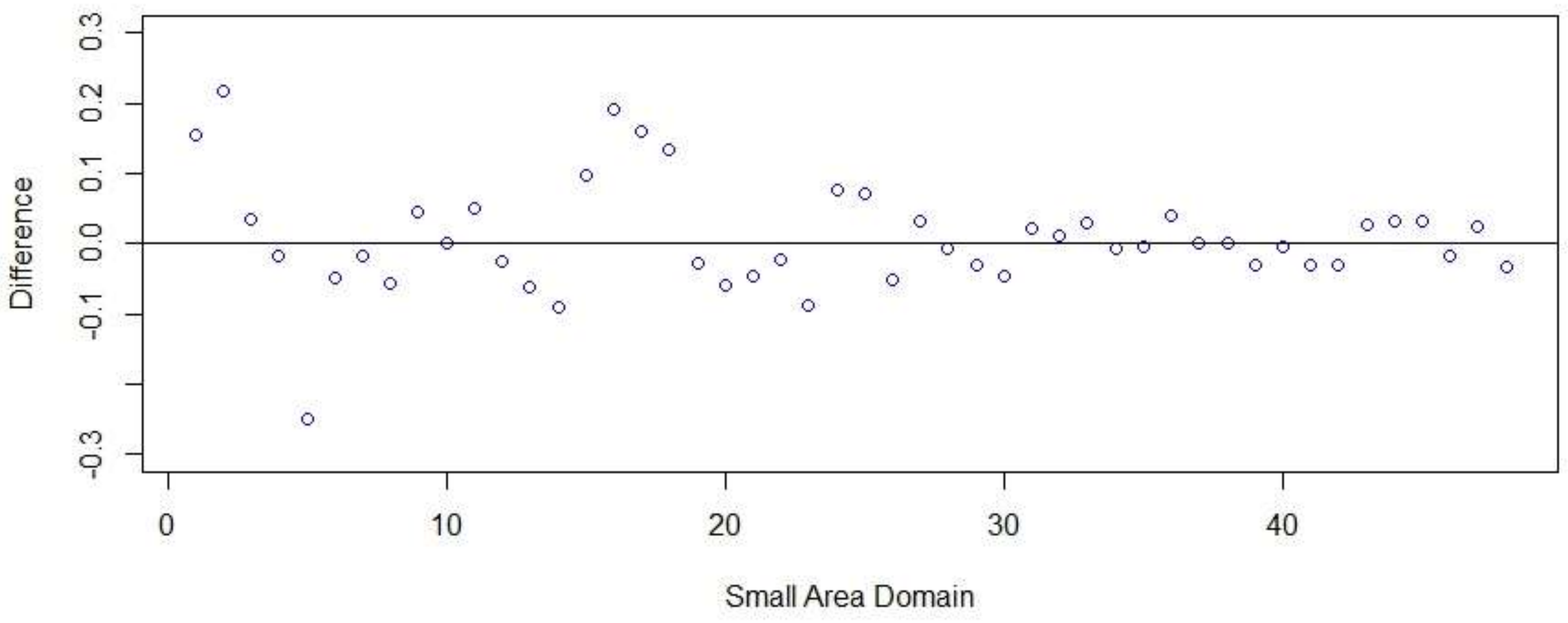


Model 1: Ratio of Standard Errors

- Ratio of standard errors for direct and model estimates for “ever had a food allergy”
- Domains are ordered based on domain sample size
- Median ratio of standard error across all domains is 2.1
- For 34 domains, the ratio of standard error was > 1.5

Model 1: Difference between Direct & Model Estimates

- Difference in direct and model estimates for “ever had a food allergy”
- Domains are ordered based on domain sample size
- Mean and median difference across all domains was approximately 0



NORC at the UNIVERSITY of CHICAGO

Bayesian approach for model 2

Using probability and non-probability survey data

- Easy to compute measure of variability of the estimates (posterior standard deviations).
- Direct estimates y_d^P from probability sample for domain *d* are unbiased

$$y_d^P = \alpha_d + x'_d\gamma + v_d + e_d^P$$

- Direct estimates y_d^{NP} from non-probability sample for domain *d* are biased

$$y_d^{NP} = \alpha_d + \beta_d + x'_d\gamma + v_d + e_d^{NP}$$

  - We assume normal prior (mean=0, variance=10^6) prior for group-level effects for race, age, gender, education α_d.
  - **Bias term $\boldsymbol{\beta_d}$** group-level effects for race, age, gender, education,
  - v_d ‘is a domain-level random effect

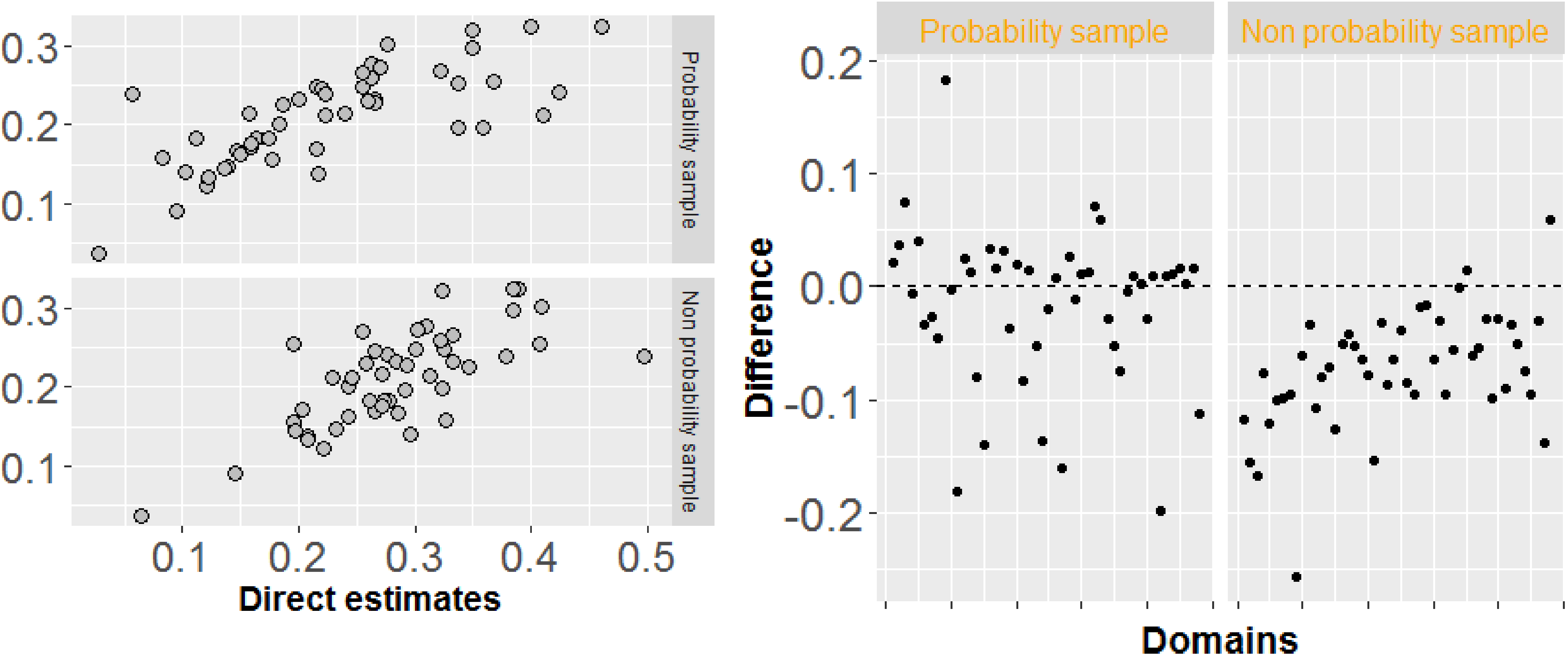
Bayesian approach

Prior distributions

- **Bias term $\boldsymbol{\beta_d}$** group-level effects for race, age, gender, education,

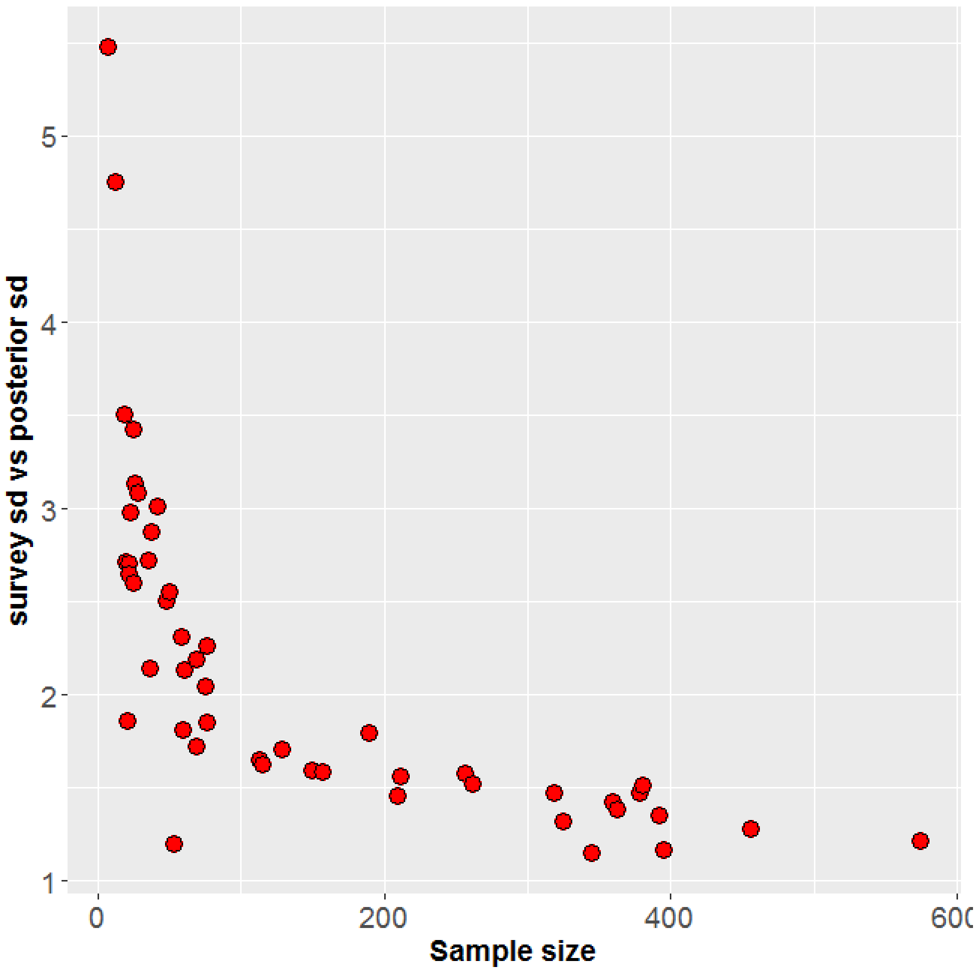
  $\boldsymbol{\beta_d}$~ N(μ_β, σ_β^2), setting μ_β=0 or alternatively μ_β ~ N(0,10^6)
- $v_{\boldsymbol{d}}$~ N(0, σ_v^2) for all 48 domains.
- Diffuse inverse-gamma priors are used for σ_β^2 and σ_v^2.
- Diffuse multivariate normal prior for γ.

Estimates (left panel), difference between model and survey estimates (right panel)



Variability of the estimates (ratio of posterior sd and survey standard error) sorted based on sample size (probability survey)

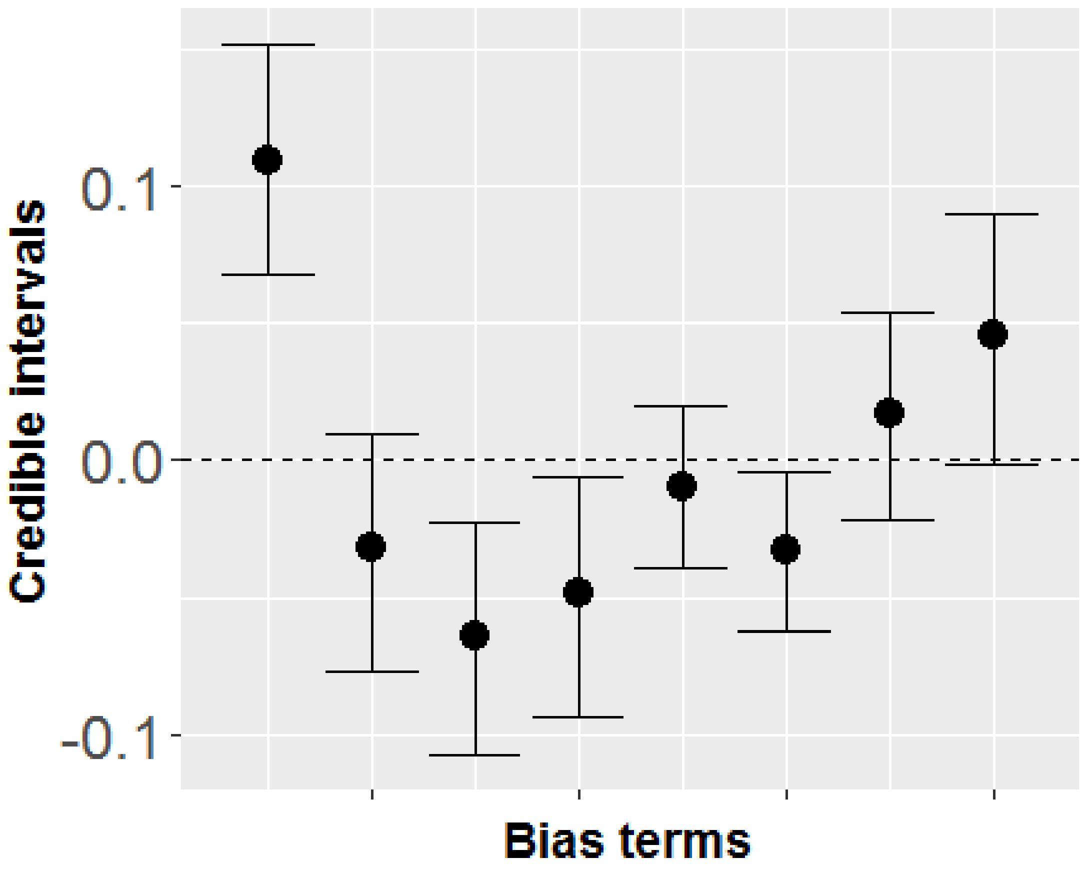
Variability of the estimates



NORC at the UNIVERSITY of CHICAGO

Bias terms

95% credible intervals



Summary and Future research

- Small area estimation models were used to combine probability and non-probability samples
- Models indicated reasonable reduction in standard error, especially for domains with smaller sample sizes

Future research and potential developments

- Auxiliary data from other sources
- Measurement error models.
- Unit-level models.

A. Chakraborty
Chakraborty-adrijo@norc.org
N. Ganesh
nada-ganesh@norc.org

Thank You!

NORC at the UNIVERSITY of CHICAGO

insight for informed decisions™