Dynamic Predictions for the Current Population Survey

Eli Kravitz

New Light Technologies, Inc.

May 11, 2023

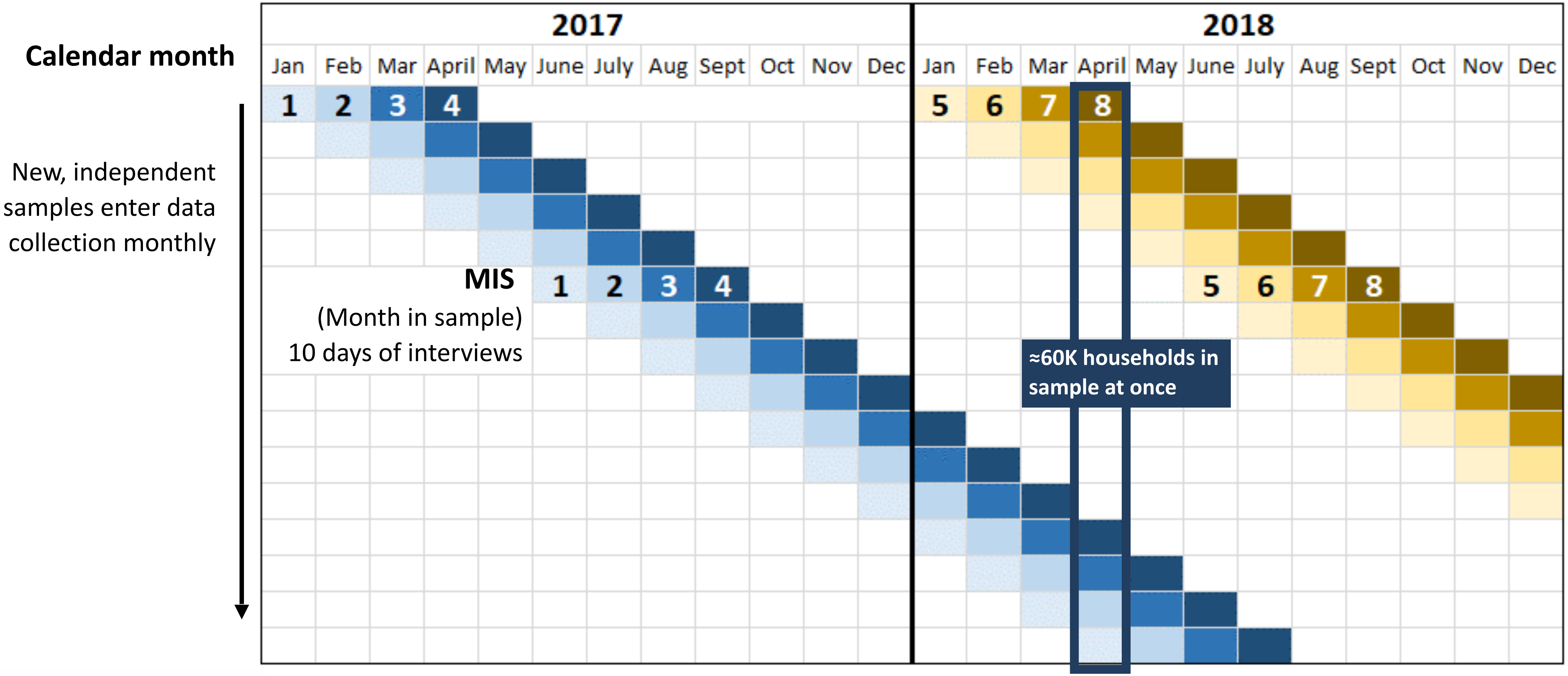
CBDRB-FY2023-CES005-018

Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau

The Current Population Survey (CPS)

- The CPS is a monthly survey of labor force participation
  - sponsored by the US Census Bureau and the Bureau of Labor Statistics
- Primary source of labor force statistic for the United States
- Unique sampling design:
  - ≈60,000 households are in sample at any given time
  - Interviewed for 4 months, out-of-sample for 8 months, then interviewed for 4 more months
  - Interviews happen during a 10 day period each month

CPS Survey Design



Calendar month

New, independent samples enter data collection monthly

MIS

(Month in sample)

10 days of interviews

≈60K households in sample at once

United States® Census Bureau

Overview of Work

- Predict whether a sampled household will respond to the Current Population Survey (CPS) during the current interview period.
    - Given no response yet, how likely is a response before day 10?

Key features of our work:

- Expand the set of covariates to include administrative records ("adrec") from other agencies and third-party data.
- Dynamic predictions
- Predict response as early as possible

Survey Paradata

MAFID	MIS	Contact attempt	Attempt outcome	MIS outcome
1	3	1	Left note at door	Refused
1	3	2	Hung up	Refused
1	4	1	Refused	Complete
1	4	2	Insuff. partial	Complete
1	4	3	Complete	Complete

- Data collected about the interview and survey process

- Includes variables like:
  - # of contact attempts
  - # of refusals
  - Did interviewer leave voicemail?
  - Responses in past months

Paradata Changes as Data is Collected

MAFID	**MIS**	Contact attempt	Attempt outcome	**MIS outcome**
1	3	1	Left note at door	**Refused**
1	3	2	Hung up	**Refused**
1	4	1	Refused	**Complete**
1	4	2	Insuff. partial	**Complete**
1	4	3	Complete	**Complete**

Goal: Will a household respond before the end of the interview period?

Paradata Changes as Data is Collected

MAFID	MIS	Contact attempt	Attempt outcome	MIS outcome
1	3	1	Left note at door	**Refused**
1	3	2	Hung up	**Refused**
1	4	1	Refused	**???**
1	4	2	Insuff. partial	**??**
1	4	3	Complete	**??**

Goal: Will a household respond before the end of the interview period?

Current Time

What is our prediction with the data we observe up to this point?

Paradata Changes as Data is Collected

MAFID	MIS	Contact attempt	Attempt outcome	MIS outcome
1	3	1	Left note at door	**Refused**
1	3	2	Hung up	**Refused**
1	4	1	Refused	**???**
1	4	2	Insuff. partial	??
1	4	3	Complete	??

Goal: Will a household respond before the end of the interview period?

Current Time

Does our prediction change?

Geographic Level Information

- Gives us information at block-group level
  - Less precise than case-level (household) information

- Data Sources:
  - **Planning Database (PDB):** Block-group demographics, responses rates to ACS
  - **Decennial:** Urban or rural indicator at block-group level
  - **Internet Access from FCC:** high-speed internet, # of internet providers

Adrec and Third Party Data

Improve predictions with data from other federal agencies and private companies:

- Tax records from IRS
- Housing information from Black Knight, Inc.
- Public assistance from HUD
- Change of address information from USPS

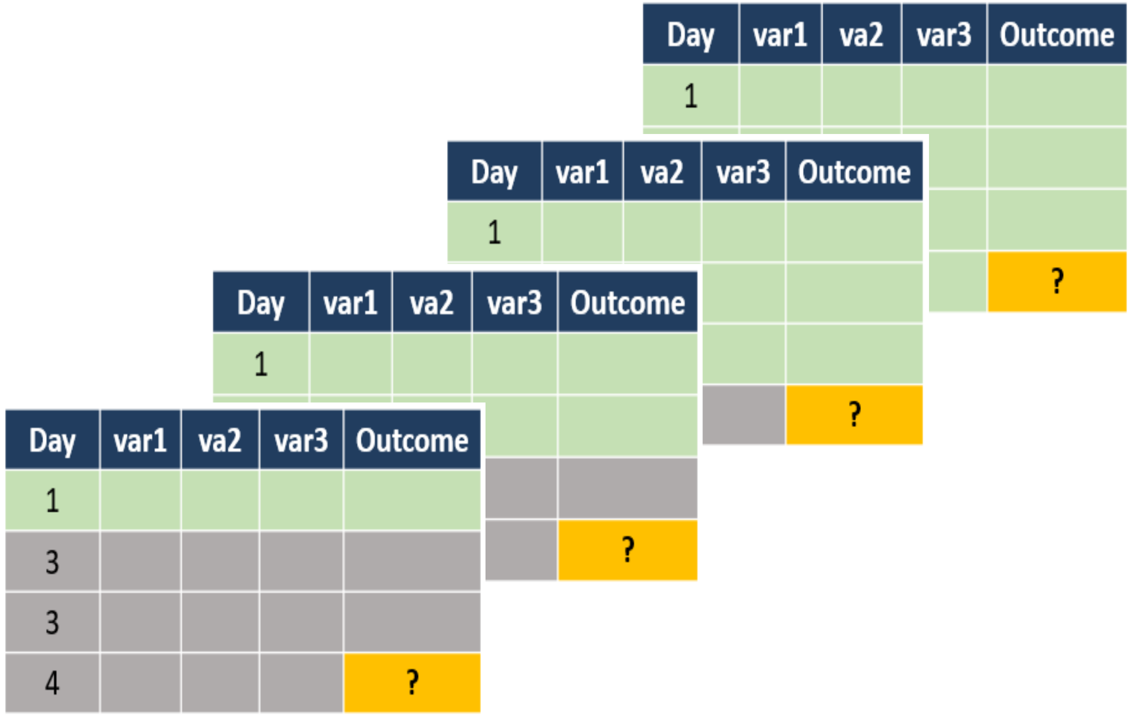
Models for Each Interview Day

- Restrict data to unresolved cases and fit separate model for each day:
  - End of Day 1: Model with accumulated paradata → predict final case resolution status

  ⋮

  - End Day 9: Model with remaining households + accumulated paradata → predict case resolution status for remaining households

- Benefits of this approach:
  - Predictions use new data as it becomes available
  - Directly estimates the quantity we're interested in:
    - Given no response by i^{th} day, how likely is a response before day 10



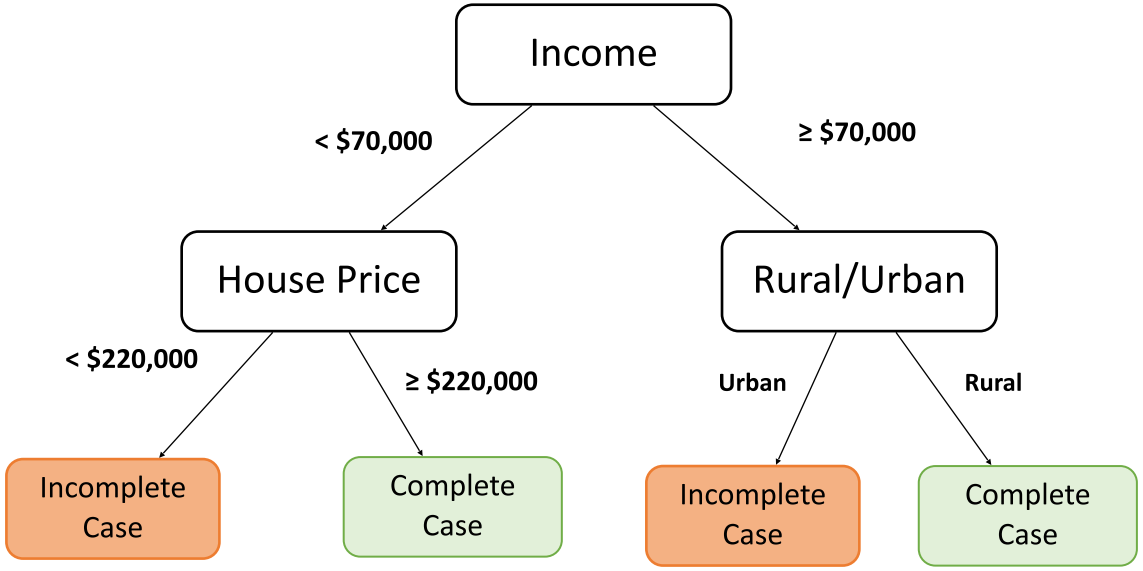
Day	var1	va2	var3	Outcome
1				
3				
3				
4				?

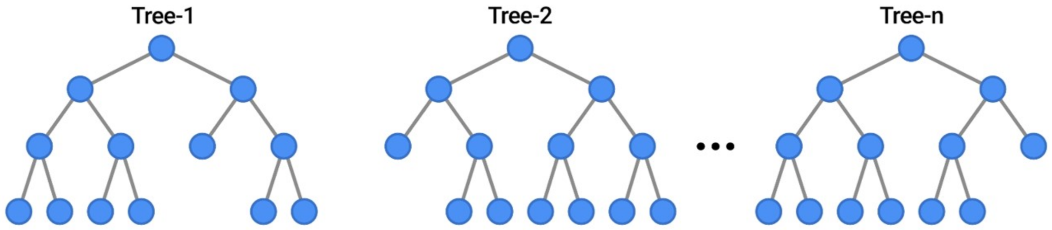
United States® Census Bureau

Modeling Approach: Tree-Based Models

- Boosted trees performed the best of all models we tried

- Tree-based Predict case response with a series of if-then rules



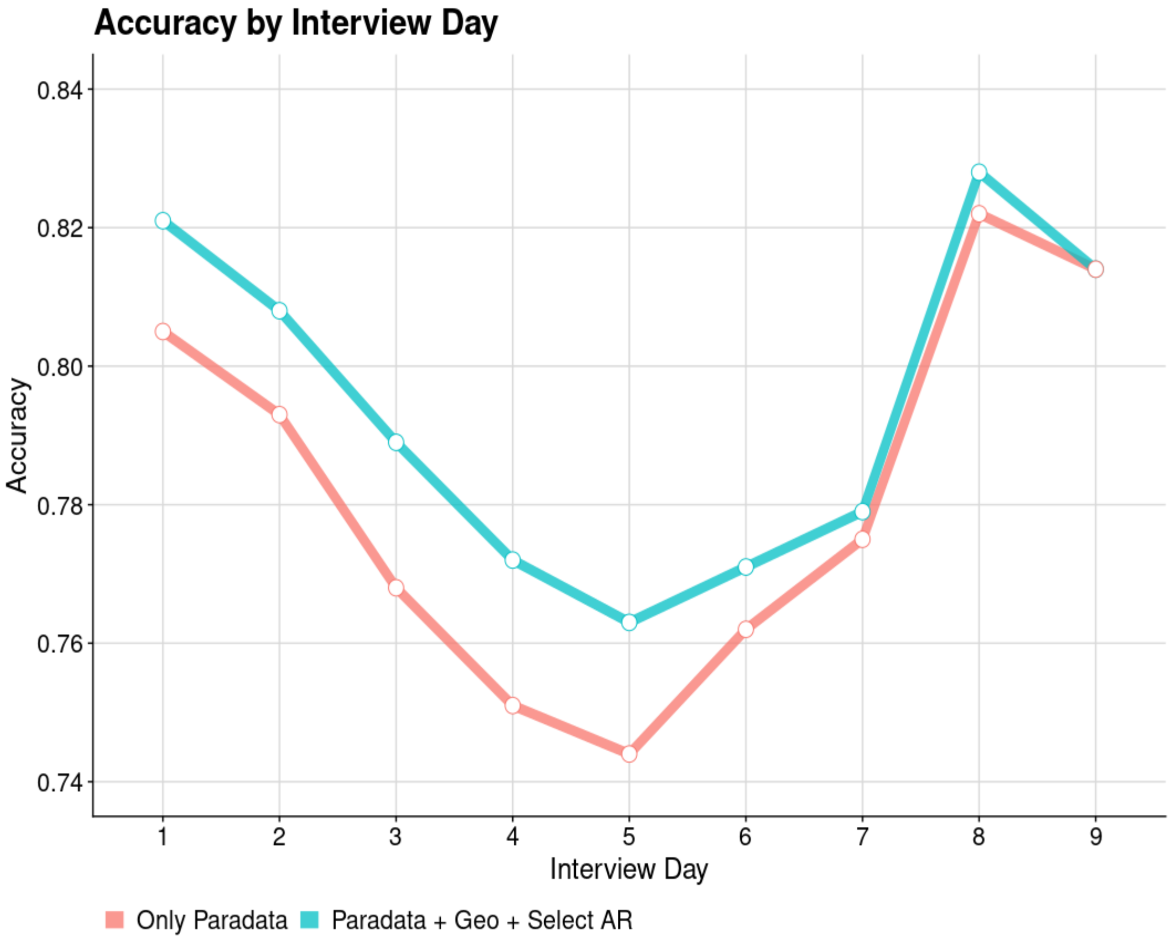


Accuracy by Day

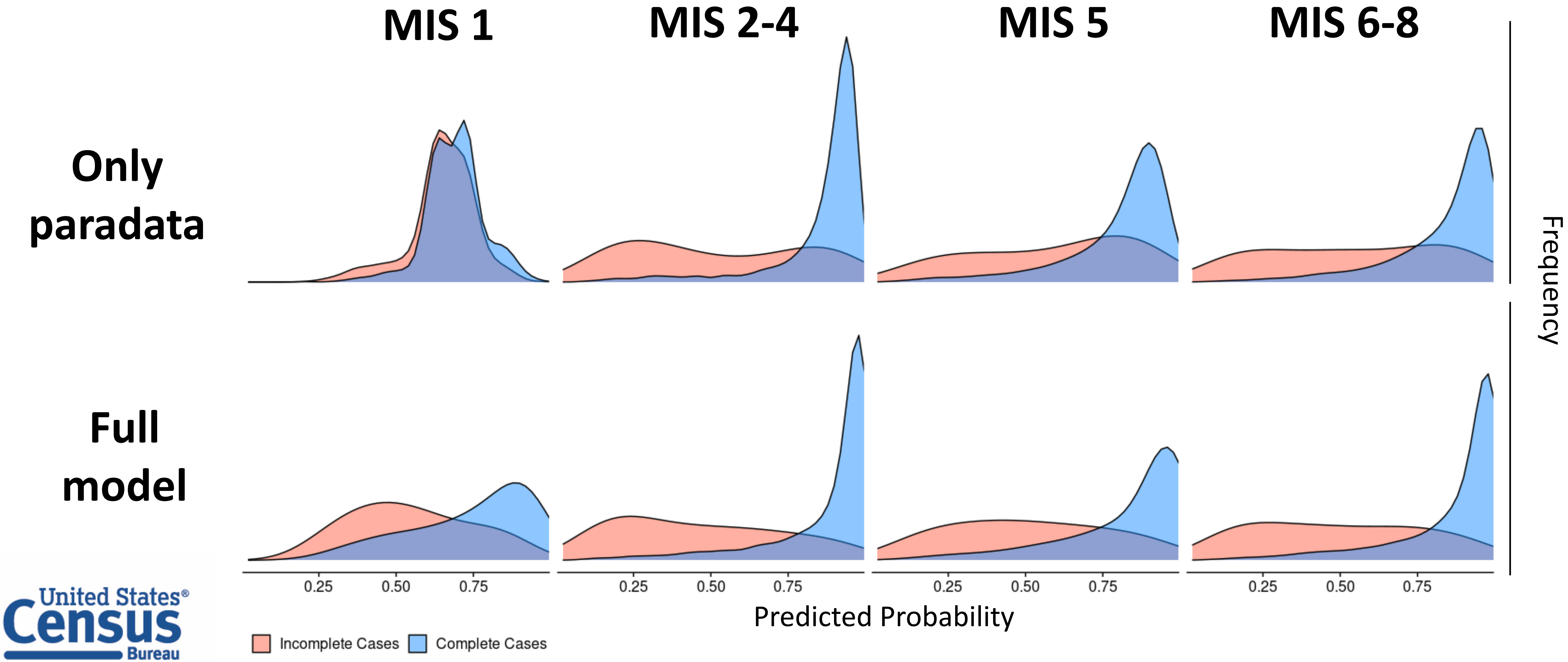
- Increase in accuracy from including adrec and geographic data.

- Still expecting further increases from adrec data.

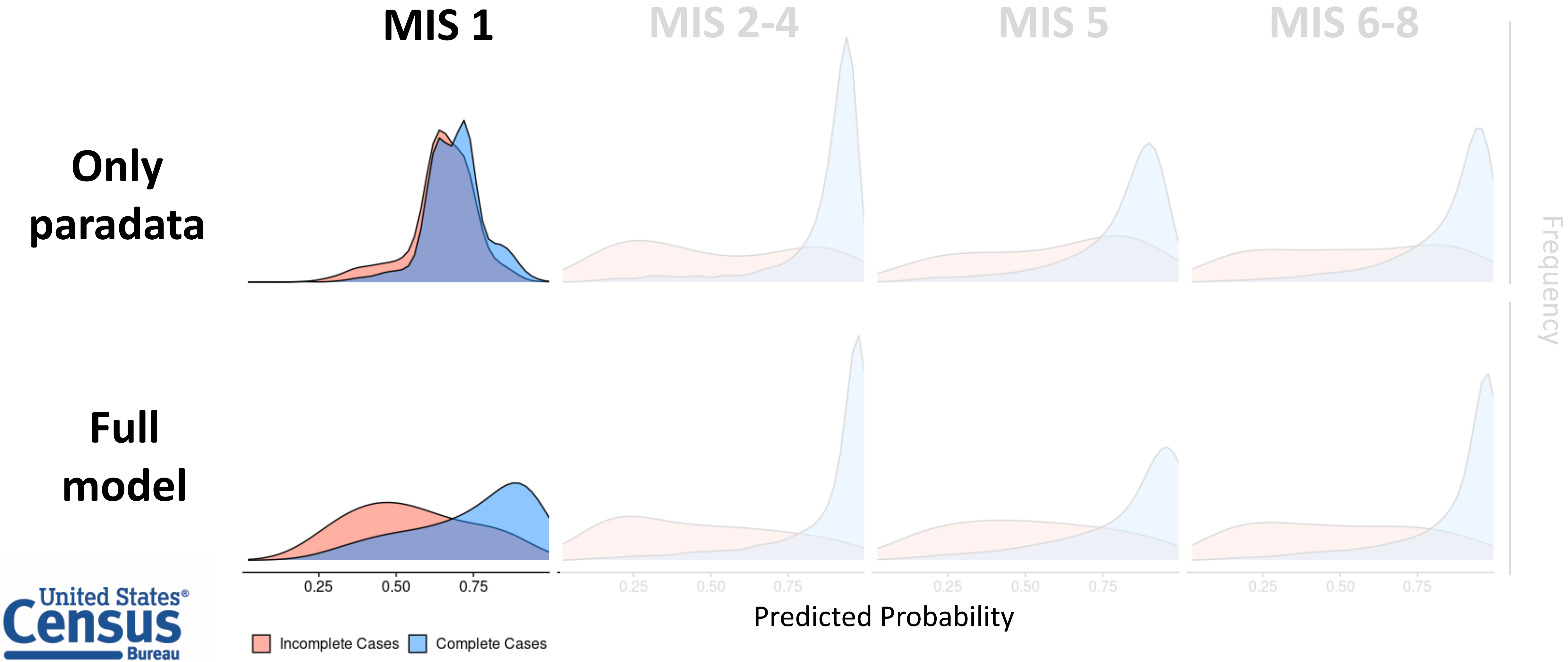
- Highest accuracy near day 1 and day 9, drop in middle in interview period



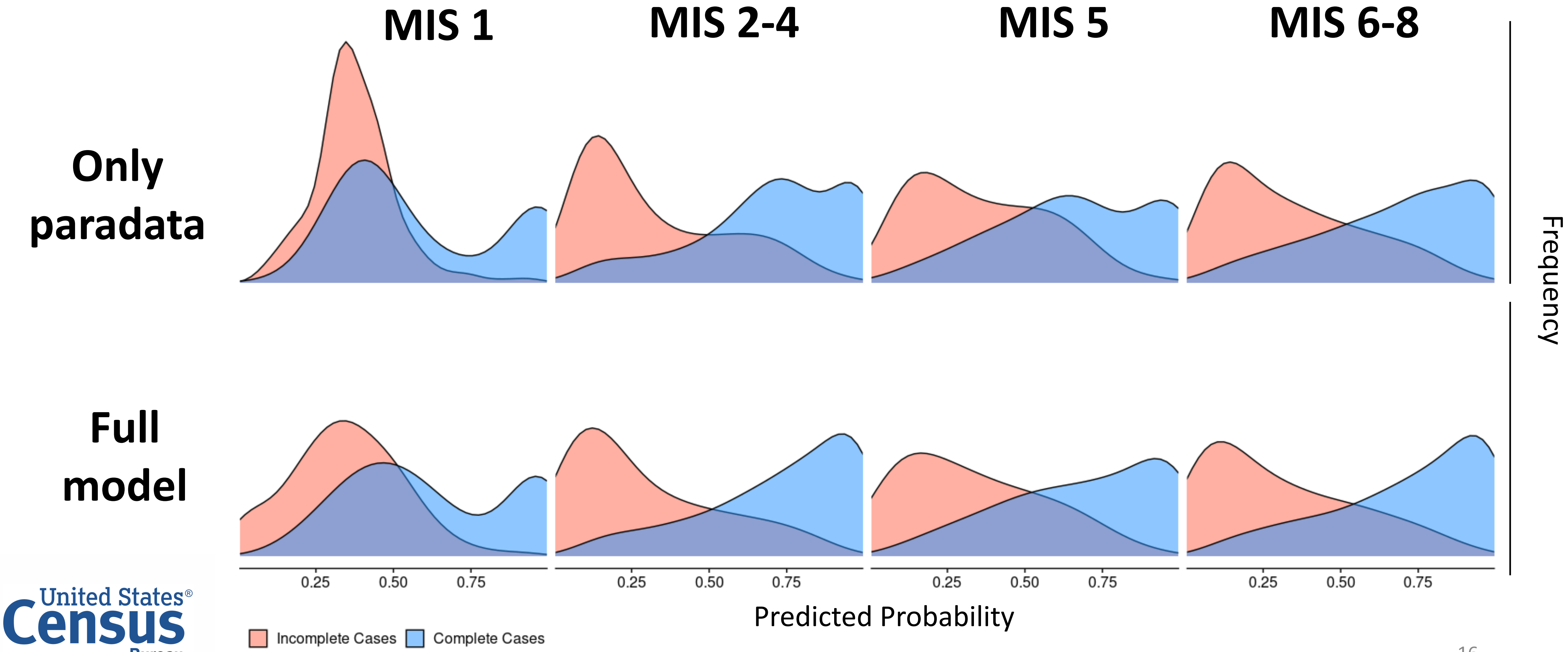
Day 1 Response/Nonresponse Separation



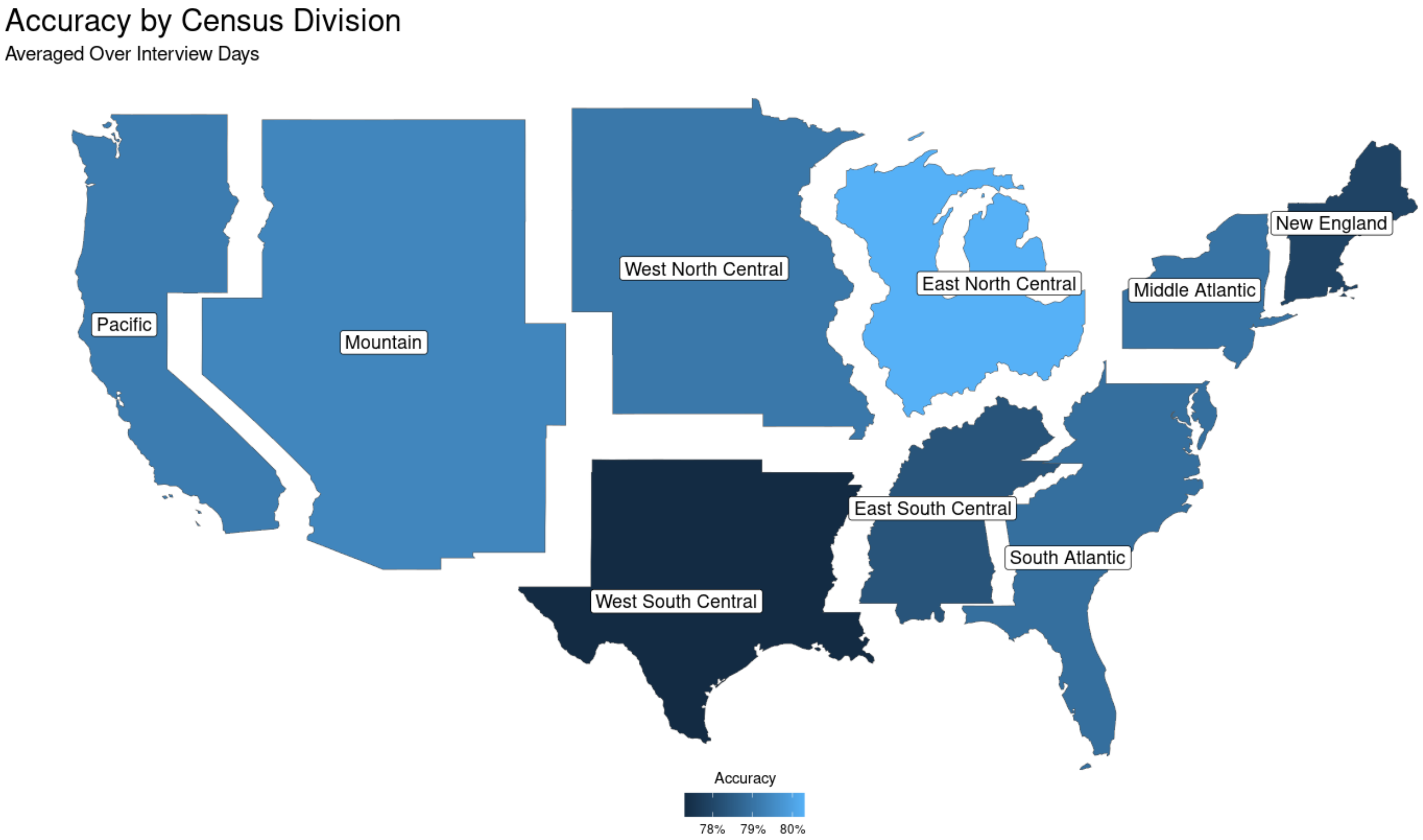
Day 1 Response/Nonresponse Separation



Day 5 Response/Nonresponse Separation



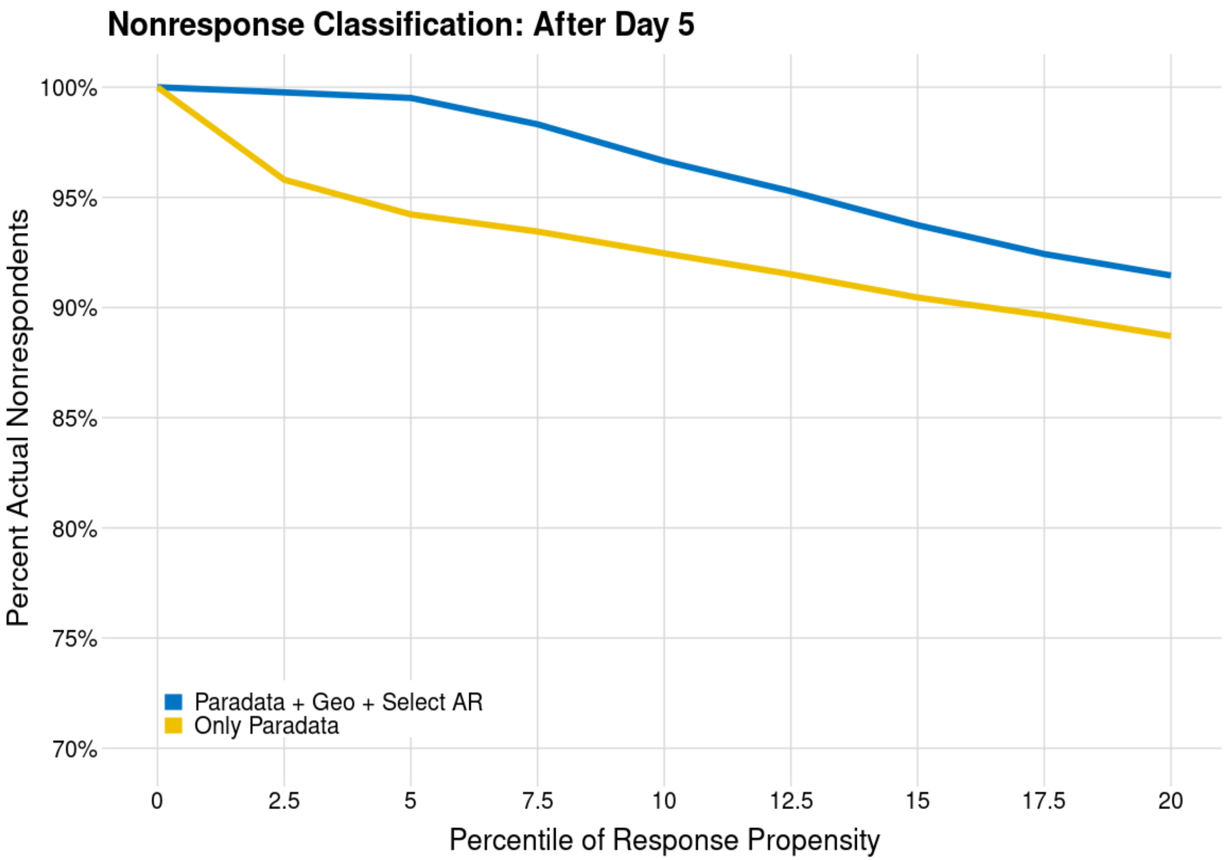
Geographic Variability



Division	Accuracy (%)
East North Central	80.3
East South Central	78.3
Middle Atlantic	79.0
Mountain	79.4
New England	77.9
Pacific	79.2
South Atlantic	78.9
West North Central	79.1
West South Central	77.3

Conceptual Use: Likely Non-Respondents

- Identify households *least* likely to complete their survey.
  - Give up on these cases or focus on a few important cases
- Ex: There are 12,000 open cases on day 5
  - Drop lowest 5% cases:
    - 600 fewer caser
    - >99% of cases won't respond
    - ≈ 6 would have responded
  - Drop lowest 10% of cases:
    - 1200 fewer cases
    - ≈97% of cases won't respond
    - ≈ 36 would have responded



Dynamic Predictions for the Current Population Survey

Eli Kravitz

May 11, 2023

CBDRB-FY2023-CES005-018

Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau